From CLINICAL NEUROSCIENCE

Karolinska Institutet, Stockholm, Sweden

# ON GENETICS AND TRANSCRIPTOMICS OF MULTIPLE SCLEROSIS

Boel Brynedal


Karolinska Institutet

Stockholm 2009

Published by Karolinska Institutet. Printed by Reproprint.



ISBN 978-91-7409-309-4

# ABSTRACT

Multiple sclerosis (MS) is a chronic inflammatory disease of the central nervous system, where both genetic and environmental factors influence one individual's risk of developing the disease. This thesis is focused on genetic and transcriptomic aspects of MS. Twelve genes have been investigated genetically for their possible independent and interaction mediated effects on MS susceptibility and clinical phenotypes, of which five genes were assessed more thoroughly: *HLA-A*, *HLA-DRB1*, *CD58*, *HDGFRP3* and *RPL5*.

The MS association with *HLA-A* was investigated in **Study I** using a cohort consisting of 1,084 MS patients and 1,347 controls. Logistic regression modelling firmly established an association suggesting a protective effect of the HLA-A*02 allele (OR: 0.63, p-value: $7x10^{-12}$).

In **Study IV** *CD58*, *HDGFRP3* and *RPL5* were investigated genetically due to previously suggested association in a genome wide association study, and because they had shown differential expression in the CSF of MS patients (**Study III**). *CD58* and *RPL5* were confirmed to be associated with MS susceptibility in 1,077 MS patients and 1,217 controls. SNPs in *CD58* conferred a multiplicative effect (ORs: 1.4-1.2, p-values: $8x10^{-5}$ – $3x10^{-2}$), whereas the effect of *RPL5* variants on MS susceptibility was conferred by the heterozygotes (OR: 1.2, p-value: $2x10^{-2}$). These genes were suggested to affect MS independently of each other as well as other known risk factors: sex, *HLA-DRB1, IL7R, IL2Rα, CLEC16A, CD226*, *SH2B3* and *KIF1B*. The interplay between these factors was elucidated, and possible epistatic effects were discovered that warrant further investigation.

Furthermore, we confirmed the association between HLA-DRB1*15 and lower age at onset, but alleles of neither *HLA-A*, *CD58*, *HDGFRP3* nor *RPL5* were found to affect severity or course of disease in **Study II & IV.**

In **Study III** gene expression profiling was performed for the first time in CSF cells from MS patients and over 4,000 transcripts were found to be differentially expressed. Simultaneously gene expression was also investigated in peripheral blood lymphocytes (PBL), and patients in an active phase of disease (relapse) were compared to those sampled in remission. These four comparisons revealed that in contrary to cells of the CSF, PBL samples did not show differential expression between MS patients and controls. Intriguingly, when comparing MS patients in relapse to those in remission, PBL samples showed more than 1,000 differentially expressed transcripts whereas in CSF cells no transcripts were differently expressed. Our results imply that MS is accompanied by active and proliferating cells in the CSF, distinguished by the regulation of genes belonging to immune related pathways. The differential expression in blood lymphocytes was characterized by a generally higher expression in relapse but with lower metabolism of several amino acids. The regulation in PBL, but not in CSF cells, implies the importance of peripheral events in driving a disease bout in MS.

# LIST OF PUBLICATIONS

I. HLA-A confers an HLA-DRB1 independent influence on the risk of multiple sclerosis. **Boel Brynedal**, Kristina Duvefelt, Gudrun Jonasdottir, Izaura M Roos, Eva Åkesson, Juni Palmgren, Jan Hillert. PLoS ONE, 2007, 2(7):e664

II. The impact of HLA-A and -DRB1 on age at onset, disease course and severity in Scandinavian multiple sclerosis patients. Cathrine Smestad, **Boel Brynedal**, Gudrun Jonasdottir, Åslaug R Lorentzen, Thomas Masterman, Eva Åkesson, Anne Spurkland, Benedicte A Lie, Elisabeth G Celius, Jan Hillert, Hanne F Harbo. European Journal of Neurology, 2007, 14(8):835-40.

III. Global expression profiling in multiple sclerosis: A disease of the central nervous system, but with relapses triggered in the periphery? **Boel Brynedal**, Mohsen Khademi, Erik Wallström, Jan Hillert, Tomas Olsson, Kristina Duvefelt. Manuscript.

IV. CD58 and RPL5 in Multiple Sclerosis: differential expression and genetic associations. **Boel Brynedal**, Izaura Lima Bomfim, Kristina Duvefelt, Jan Hillert. Manuscript.

# CONTENTS

# LIST OF ABBREVIATIONS

| | |
|---|---|
| BBB | Blood Brain Barrier |
| CI | Confidence Interval |
| CNS | Central Nervous System |
| CSF | Cerebrospinal Fluid |
| EAE | Experimental Autoimmune Encephalomyelitis |
| EDSS | Expanded Disability Status Scale |
| EM | Expectation-Maximization |
| FDR | False Discovery Rate |
| GWAS | Genome Wide Association Study |
| HGNC | HUGO Gene Nomenclature Committee |
| HLA | Human Leukocyte Antigen |
| HWE | Hardy-Weinberg Equilibrium |
| IMSGC | International MS Genetics Consortium |
| IPA | Ingenuity Pathway Analysis |
| KEGG | Kyoto Encyclopedia of Genes and Genomes |
| LD | Linkage Disequilibrium |
| MALDI-TOF MS | Matrix Assisted Laser Desorption Ionisation Time-Of-Flight Mass Spectrometry |
| MM | Mismatch |
| MRI | Magnetic Resonance Imaging |
| MS | Multiple Sclerosis |
| MSSS | Multiple Sclerosis Severity Scale |
| NK | Natural Killer |
| OND | Other Neurological Diseases |
| OR | Odds Ratio |
| PBL | Peripheral Blood Lymphocytes |
| PCA | Principal Component Analysis |
| PCR | Polymerase Chain Reaction |
| PM | Perfect Match |
| PP | Primary Progressive |
| QC | Quality Control |
| qRT-PCR | Quantitative Real-Time PCR |
| RIN | RNA Integrity Number |
| RMA | Robust Multi-array Average |
| RR | Relapsing Remitting |
| SNP | Single Nucleotide Polymorphism |
| TAP | Transporter, ATP-binding Cassette |
| TCR | T-Cell Receptor |

# 1 BOEL'S THESIS

All studies included in this thesis focus on Multiple sclerosis (MS). Typically we, as genetic researchers, describe MS as a complex disease meaning that several genetic and environmental factors influence an individual's risk of developing the disease. These genetic risk factors have been pursued ever since it was shown that, in fact, inherited factors play a role in MS (reviewed in [1]). As early as in the 1970[ies] the first genetic factor affecting MS susceptibility was discovered through functional tests. After some discussion, it was established that a variant of the class II region of the major histocompatibility complex, also called the human leukocyte antigen (HLA) class II region, conferred a risk of developing MS [2]. In the past few years a lot has been accomplished, and several genetic risk factors have been identified. Major collaborative projects as well as the technical and methodological progress promise to move this field of research ever further. The genetic studies included in this thesis play a small role in this development, and these as well as theoretical and methodological aspects of genetic studies are discussed in chapter three.

Processes are ongoing within MS patients, dependent or independent of present genetic variants, which we would like to elucidate, characterize and understand. This thesis includes a study investigating the transcriptome of two tissues from MS patients: cerebrospinal fluid (CSF) cells from the central nervous system (CNS), which has never been assessed for gene expression profiling earlier and peripheral blood lymphocytes (PBL). Gene expression profiling and the results from our investigation are discussed in chapter four.

# 2 MULTIPLE SCLEROSIS

## 2.1 CLINICAL ASPECTS

There are no specific tests that confirm a diagnosis of MS, rather MS is in principle diagnosed on clinical grounds when there is evidence of lesions in the central nervous system (CNS) in the form of at least two clinical bouts of neurological symptoms affecting at least two anatomic sites of the CNS, and no better explanation for the clinical and paraclinical abnormalities exist. Even a single neurological bout can sometimes allow a diagnosis of MS if it is accompanied by the observation of a subsequent lesion by magnetic resonance imaging (MRI) or abnormal evoked potentials. Thus, the patient history or/and laboratory investigations, such as cerebrospinal fluid (CSF) analysis, and radiological investigations, such as evoked potentials and MRI, should indicate dissemination in time and space [3].

The neurological symptoms vary considerably between and even within an individual patient, but often reflect the location of the lesion within the CNS. Symptoms include, but are not limited to, fatigue, numbness, muscular weakness, balance problem, blurry vision, pain, bladder dysfunction and cognitive impairment. Typically the disease bouts arise over hours or days, then plateau and improve (sometimes incompletely) over days to weeks. Many lesions are clinically silent, in fact, MRI studies have indicated that lesions appear seven to ten times more frequent than clinical relapses [4]. The number of lesions correlates rather poorly with disability, whereas measures of brain atrophy possess better correlation [5].

Disability for individuals with MS is most commonly assessed using are the expanded disability status scale (EDSS) developed by Kurtzke [6]. Here much of the complexity of the condition is assessed through both impairment and disability. The EDSS has shown low inter-rater reliability, and this ordinal scale is not evenly distributed [7]. The EDSS is a measurement of the experienced disability in one individual at a given time point, and does not consider disease duration, thus the severity of disease course for two individuals is difficult to compare. Several methods have been proposed to assess severity; the multiples sclerosis severity scale (MSSS) uses EDSS in conjunction with disease duration, and an individual is compared to others in a large longitudinal database and thus assesses cross-sectional disability [8]. Thus, acquiring an MSSS

under 1 signifies that ones disability is among the lowest 10 percent as compared to the global MSSS. Another approach to assess severity includes survival analysis where e.g. time to EDSS 6 can be analysed, this will be discussed in later sections.

Initially, the disease course is usually (in most studies 80-90 %) characterized by relapses (disease bouts) and remission (periods of recovery), called a relapsing remitting form of MS (RRMS). A majority of these patients develop gradual progression between relapses after a median duration of 19 years, and then entered the secondary progressive (SP) phase of the disease [9]. For a lesser proportion (5-20 %) the initial disease course is primary progressive (PPMS), characterised by a steady progression from onset. The majority of MS patients are female, with a female to male ratio of approximately 2.5 in our patient cohort, although the PP course is more common among men [9].

There is no cure for MS today, but several disease modifying treatments exist that reduce the number of clinical relapses and lesions seen by MRI. Different Interferon β preparations (Betaferon, Rebif and Avonex) and glatiramer acetate (Copaxone) have been used starting in the mid nineties, and more recently several monoclonal antibody based medications as well as oral immunomodulatory treatments have appeared and are currently under investigation [10].

## 2.2 IMMUNOLOGY AND PATHOLOGY

MS is an inflammatory disease where the leukocytes, for unknown reasons, attack the oligodendrocytes that produce the myelin surrounding the axons in the CNS. Myelin is made up by multiple layers of cellular membrane arranged in segments along the axons enabling the saltatory conduction of action potentials. The inflammation of myelin is accompanied by demyelination and neurodegeneration, and remyelination is present to some extent. Lesions with ongoing active inflammation are associated with damage to the blood brain barrier (BBB), indicating an increased influx of immune cells into the CNS. Thus, these lesions are usually oriented around blood vessels, and characterized by infiltration of lymphocytes and macrophages, the loss of myelin sheets and axons that are embedded in astroglial scar tissue [11]. The majority of infiltrating lymphocytes are T-cells, and include both CD4+ T-cells – restricted to activation by HLA class II presenting cells, and CD8+ T-cells – restricted to activation by

HLA class I presenting cells. CD8+ T-cells have been shown to bind to oligodendrocytes and axons in MS lesions and the T-cell infiltration has been associated to HLA class I expression, thus CD8+ T-cells may play an important part in the pathology (reviewed in [11]), although MS has long been viewed as a CD4+ T-cell mediated disease [12,13]. Both T-cells (CD4+ as well as CD8+) and B-cells undergo clonal expansion within the CNS. Besides the sharply demarcated lesions there is also a more diffuse change (signal alterations) and infiltrates of cells in normally appearing white matter and a brain atrophy that is sometimes profound in latter stages of disease [11,14]. Whether the inflammation or the neurodegeneration is the primary cause of MS symptoms, and whether neurodegeneration can occur in MS independent of inflammation has been debated [15,16].

One major hypothesis of the initiation of MS is that (autoreactive) myelin specific T-cells are activated in the periphery and migrate to the CNS. The activation could be facilitated by e.g. molecular mimicry [17] or incorporated myelin proteins in viruses. Once across the BBB, the T-cell gets reactivated and release pro-inflammatory molecules that further facilitate the recruitment of immune cells into the CNS. Autoimmunity is present in anyone to some extent and is usually harmless, but can also cause autoimmune diseases. The autoimmune, or immunological, response in MS is a likely cause of most, if not all major symptoms, and the reason for its initiation is of vital importance.

### 2.3 EPIDEMIOLOGY

MS has an uneven distribution throughout the world, with lower prevalence closer to the equator [18], and generally prevalence is highest in northern Europe, southern Australia and north America. Accordingly, Scandinavia has a high prevalence of about 1 in 1000 individuals. Most indigenous people, such as the Sami population of Scandinavia, do not follow this pattern and show a lower prevalence of MS. It has been proposed that the latitude gradient is caused by the emigration of northern European individuals to regions with similar climate, thus indicating that northern European genetic variants increase the risk of MS, and the lower prevalence of MS among indigenous, more isolated, populations could reflect that as well [1]. There are however migration studies that indicate environmental influence that cannot be explainable by genetics, these studies are often small and individual results might be questionable. As

discussed by Compston and Confavreux [18], migration from high risk areas to low risk areas is suggested to decrease ones risk of developing disease, *if* individuals were below fifteen years of age, and migration from low risk areas to high risk areas has shown the corresponding pattern.

The latitude gradient has also been suggested to show the environmental influence to MS susceptibility, and has been suggested to be caused by less exposure to sunlight and thereby decreased levels of vitamin D [19]. Another popular hypothesis is that infection could cause the gradient, and both a hygiene hypothesis, where early infections prevent development of MS, and hypotheses regarding specific infections that would cause MS have been postulated (reviewed in [20]). There are reports indicating that the latitude gradient is disappearing [21,22], which might be explained by an improved hygiene or spreading of certain pathogens. Alternatively, the latitude gradient might reflect that genetic factors are important in MS susceptibility, and the disappearance of this gradient could be due to increased migration throughout the world during the latest decades. The evidence for a genetic predisposition in MS is however quite robust [1], and is presented in Figure 1. Here it is shown that concurrence in MS correlates with degree of genetic sharing, with the largest concordance rates seen in monozygotic twins. Different studies have shown variations in the concordance rate of all levels, but the correlation between genetic sharing and concurrence is almost always seen.

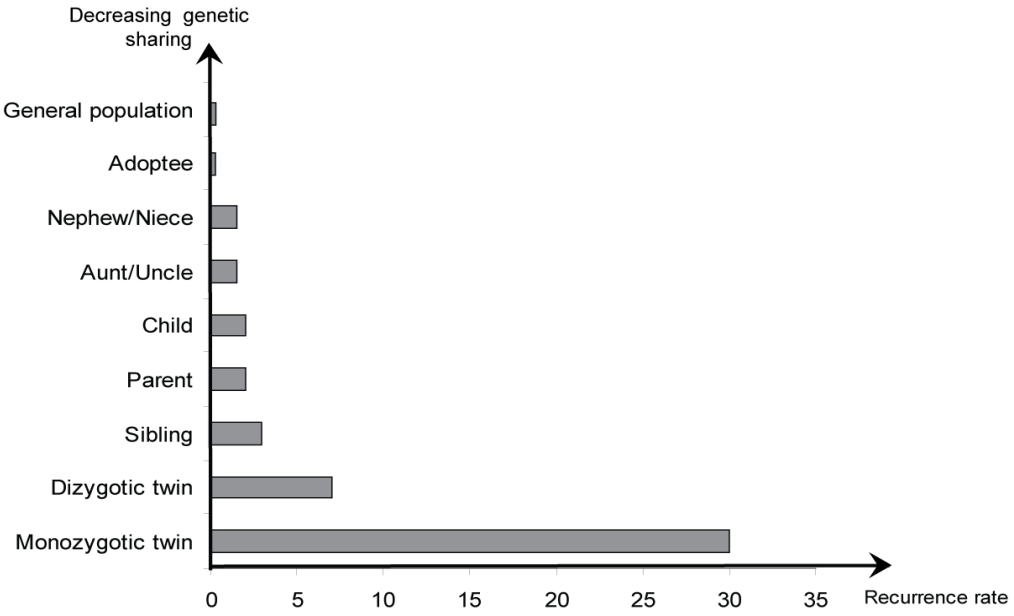


*Figure 1. The age adjusted recurrence rate of multiple sclerosis for individuals with different degrees of genetic sharing. Data from meta analysis reported in McAlpine's Multiple Sclerosis [1].*

# 3 GENETIC STUDIES

Once the existence of genetic risk factors has been demonstrated, the search for attributable variants begins. Currently, most of the genes in the human genome are known, and through resources such as the HapMap [23] database many of the polymorphisms are known as well. The daunting tasks of selecting suitable genes to investigate, the genotyping, the statistical analysis and finally functional analysis remain, and are henceforth discussed in relation to the thesis.

## 3.1 AIM

The aim of *Study I & II* was to investigate the proposed role of the HLA class I gene *HLA-A* in MS susceptibility, and how *HLA-A* and *HLA-DRB1* interact in affecting MS susceptibility and clinical phenotypes, such as age at onset, severity and disease course.

The aim of *Study IV* was to test the hypothesis that a few selected genes, having shown differential expression in the cerebrospinal fluid (CSF) of MS patients compared to controls, as well as suggestive association in a genetic screen performed by the international MS genetics consortium (IMSGC) are indeed of importance for MS susceptibility and clinical phenotypes.

## 3.2 COMPLEX GENETIC DISEASES

Complex diseases, as opposed to Mendelian diseases, do not show a clear inheritance pattern in families. Individuals with greater genetic sharing do however show higher concordance, as indicated for MS in Figure 1. Complex, or multifactorial, diseases are caused by both environmental and genetic risk factors, where one individual needs to acquire several risk factors in certain combinations, in order to develop disease. A useful simile has been provided by Rothman [24] where a sufficient cause of disease is described as a pie. Here each piece of the pie is a risk factor and disease only develops in a person who has acquired an entire pie, and moreover a disease can be caused by several different pies (see Figure 2).

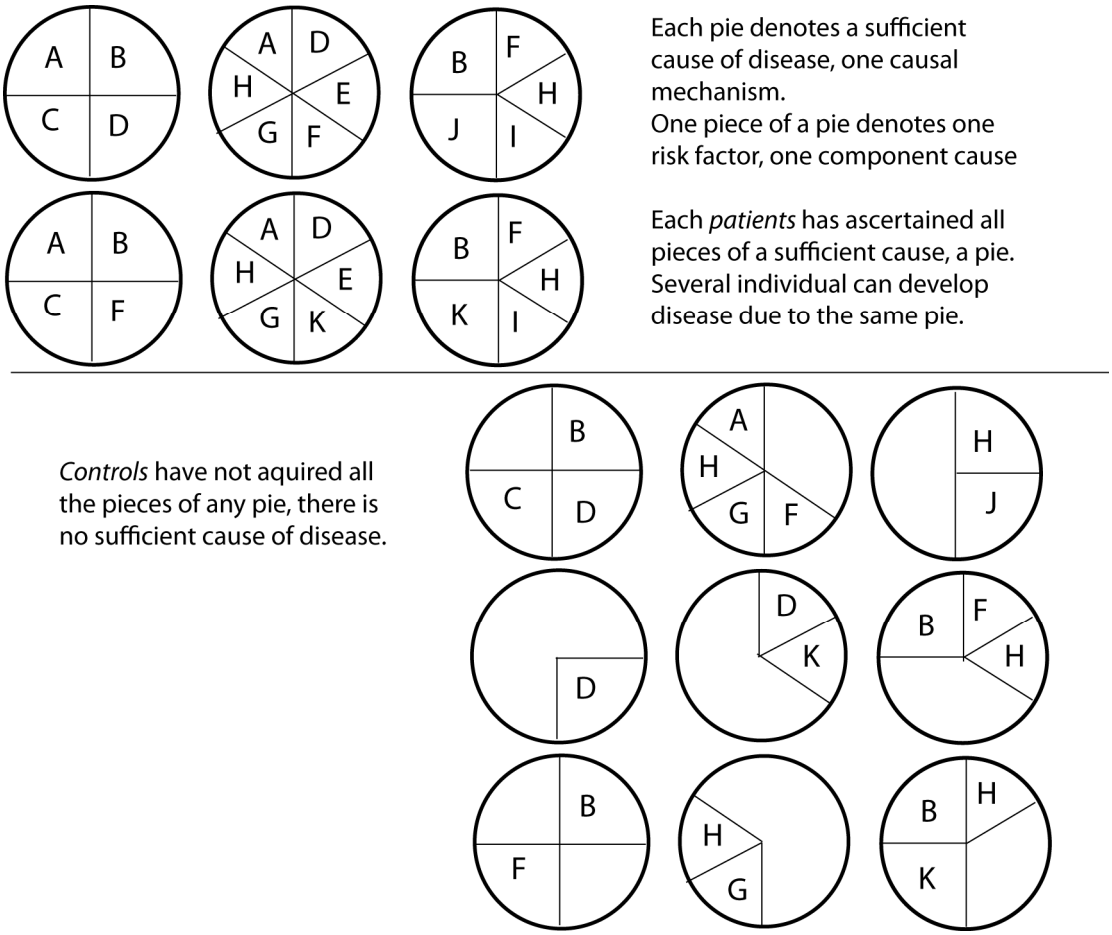


*Figure 2. Illustration of Rothmans pie model showing sufficient causes of disease (at the top) among patients, and how the risk factors, pieces of pie (here denoted with different letters), can be distributed among controls (bottom).*

Thus, a piece of a pie can be present in both patients and healthy individuals, and does not have to be involved in the development of disease in all patients. This model assumes genetic heterogeneity; a disease can develop due to several different combinations of risk factors and thus two individuals can develop disease for completely distinct reasons. Using Rothman's pie models, the strength of one component cause relates to how many of the patients of a given disease that are attributable to a pie including the component cause. Several measures are used to describe this strength; some use the frequency of the risk factor alone, others use both the frequency as well as an effect measure of the increased risk attributable to the factor in question.

In order to find genetic variants connected to disease one must find means to detect the sequence difference or variants linked to the disease causing variants. Currently techniques are evolving rapidly - individual genomes have been resequenced, and

even though this level of resolution is not yet feasible to a common laboratory, it will probably be in a few years. Additionally, epigenetic factors such as methylation, which has not yet been investigated in MS, are likely to be of importance for MS susceptibility. So far, microsatellites and single nucleotide polymorphisms (SNPs) have been the markers of choice in genetic studies rather than resequencing. Microsatellites are tandemly repeated DNA sequences with a high mutation rate, which makes them very informative since they tend to differ from person to person. SNPs are much more common than microsatellites and throughout the genome it is estimated that about 13 million SNPs exist. A single base alteration is defined as a SNP if it has a frequency of at least 1 % in the population.

## 3.3 ASSOCIATION STUDIES

In association studies individuals with the disease and controls from the population are used to find association between genotype and phenotype. A genetic variant which differs in frequency between patients and controls might be or reflect a piece of one or more disease pies.

An association study can be conducted at any level, using single markers, entire genes or genome wide association studies (GWAS). The latter approach usually includes the utilization of immobilised oligonucleotides on arrays to detect SNP alleles.

### 3.3.1 Choice of genes and markers

The included genetic studies in this thesis are candidate gene studies, where genes were selected for genetic assessment based on prior knowledge. Such information could include data from animal model studies, where a genetic variant influences the phenotype of disease, or information regarding the function of a certain gene that suits what is known about the disease aetiology. Genes with association in another disease with some phenotypic feature resembling the disease of interest, or genes with previously suggestive association could also be reasonable candidates.

Once decided which genes to investigate, genetic markers within and/or close to those genes should be evaluated. One can consider that markers in close proximity are often inherited together, that some SNPs are within coding regions, others cause amino acid changes in the resulting protein, and some might have been investigated previously.

Using data from the HapMap consortium [23] information about known SNPs and the correlation between these SNPs in eleven different populations can be extracted. If two SNPs are highly correlated it may be possible to only to genotype one of them and detect possible association, this approach is referred to as tagging. Investigating genetic association using common SNPs and a tagging approach assumes that the disease is caused by common genetic variants, or possibly rare variants that correlate with a genotyped variant.

#### *3.3.1.1 Early studies in MS (1970-2006)*

During this time HLA variants and blood groups were among the only genetic markers available to researchers, and surprisingly HLA was found to behave differently among MS patients. Genetic methodologies were not available, and HLA "phenotypes" were tested using functional tests such as the mixed lymphocyte culture reaction, microdroplet lymphocyte cytotoxicity test and/or complement fixation with platelet antibodies [25,26]. Thereby functional alleles of HLA genes were determined, and this nomenclature is still in use today.

Several HLA class I alleles were reported as associated, and somewhat later also class II alleles, and eventually HLA-DRB1*15 was shown to exhibit the strongest association to MS, and the initial class I associations were regarded as secondary [27]. MS is associated with a combination, haplotype, of HLA class II alleles: DRB1*1501, DRB5*0101, DQA1*0102, DQB1*0602 [28]. The LD within this haplotype is high, and one therefore usually genotypes only the DRB1 locus. Moreover, the majority of DRB1*15 alleles are in fact of the subtype DRB1*1501 and thus one only genotype individuals at this resolution (two digits), which also represents the resolution of the earlier functional tests. The DRB1*15 allele is present in about 55-60 % of MS patients and 30 % of healthy individuals in our data, and is associated with MS in practically all populations studied. Due to putative roles of HLA class I molecules in an animal model of MS, experimental autoimmune encephalomyelitis (EAE), and the initial class I associations to MS, HLA class I molecules were re-investigated in relation to MS susceptibility within our group more recently [29]. Here both the *HLA-A*, *-B* and *–C* genes were genotyped in 87 MS patients and 102 controls, and those alleles showing signs of association were genotyped in an additionally cohort of 113 MS patients and 108 controls. HLA-A*03, A*02 as well as B*07 were associated with MS, but HLA-B*07 was dependent on the

HLA-DRB1*15 association as shown by stratified analysis. In the stratified analysis (DR15 positive and negative groups) HLA-A*02 was suggested to be associated in both groups (p-values 0.012 and 0.06), whereas HLA-A*03 was less associated (p-values 0.072 and 0.12). These groups were apparently small, and reported p-values were corrected for multiple testing. Furthermore, the HLA-A*03 allele was found to modulate the HLA-DRB1*15 association in a Norwegian cohort, whereas independent effects of HLA-A*02 was not investigated [30]. Thus, the roles of *HLA-A* alleles in MS were not elucidated.

Prior to year 2007 hundreds of non-HLA candidate gene studies have been conducted to find genetic variants predisposing to MS. None of these provided sufficient evidence for any genetic variant to be regarded as truly associated to MS susceptibility by the MS research community. In retrospect we can conclude that our expectation regarding the effect sizes were exaggerated, and thus numerically too limited study populations were utilized. Several published studies reported significant associations, but since follow up studies failed to verify them, they were regarded as false positives. A multitude of studies showed non-significant findings, and one might assume that many more such studies failed to be published (publication bias). In my mind, these genes cannot be dismissed; there might be false negatives as well as false positives.

#### *3.3.1.2 Finally some success! (2007 and onwards)*

In the early 2000$^{th}$ the pessimism within the MS field was large due to the huge amount of "failed" candidate gene studies and the failure of a large microsatellite screen in families from several populations [31]. At the time several companies had developed techniques to detect many thousands of SNP alleles on arrays and the research community was eager to exploit their possibilities. The technique was however still expensive, and researchers had realised that the study populations needed to be larger than any individual research group usually have access to. Consequently the first GWAS in common diseases [32] and MS [33] were published during 2007 and were large collaborative projects, including 1000-2000 patients and as many controls. Prior to this the first non-HLA gene (*IL7R,* interleukin 7 receptor) had been confirmed in MS, by Lundmark et al. [34] in parallel with Gregory et al [35], when validating a smaller initial study performed a few years earlier [36]. Several studies have

thereafter validated the *IL7R* association [33,37], and the IMSGC further validated their own initial *IL2Rα* association. This report was followed by several confirmatory studies, and thereby *CD58*, *RPL5* (ribosomal protein L5) and *CLEC16A* (C-type lectin domain family 16) [38-40] were validated as associated to MS. Recently a GWAS initiated in a small Dutch isolated population indicated a role for *KIF1B* (kinesin family member 1B) in MS susceptibility, which was validated in a cohort of Canadian, Swedish and Dutch case-control populations [41]. Another GWAS performed in an exceptionally well-characterized MS population was recently published, where genetic variants were connected to susceptibility, age of onset, disease severity, brain lesion load and normalized brain volume [42]. Here *GPC5* (glypican 5) is suggested to confer risk of MS, a finding not yet validated by any other group. Genes known to be connected to type I Diabetes have also been investigated in MS, and thereby *CD226* and *SH2B3* (SH2B adaptor protein 3) were indicated as MS susceptibility genes [43,44].

#### *3.3.1.3 Choice of genes and markers in Study I & II*

As described above, a possible impact of the class I gene *HLA-A* in MS susceptibility had been reported by our group as well as others [29,30], but these studies were small, underpowered and results were unclear. Therefore we set out to investigate *HLA-A* in an independent study population with emphasis on describing the relationship between *HLA-A* and *HLA-DRB1* alleles (*Study I*) in affecting MS susceptibility and clinical variables (*Study II*).

#### *3.3.1.4 Choice of genes and markers in Study IV*

In *Study IV* genes which had shown involvement in MS both genetically and functionally were investigated. At the time of study design the IMSGC article [33] had been published and the initial results from our gene expression profiling project (*Study III*) were examined for the first time. Therefore we matched genes from all significantly differently expressed probe sets in the CSF of MS patients compared to controls, with the genes of the 110 SNPs published in the supplementary of the IMSGC paper. Six genes were suggested to be connected to MS in both studies, and three were selected for further investigation (*CD58*, *RPL5* and *HDGFRP3* (hepatoma-derived growth factor, related protein 3)), the others were dropped due to financial restrictions.

We included all SNPs within 10 kb upstream and 10 kb downstream of each gene, and selected SNPs to detect all common variation within this region. The indicated SNPs in the IMSGC study were always included, and SNPs in coding regions were included if the HapMap [23] data showed evidence of polymorphism with a frequency of more than 5 % in the CEU (CEPH (Centre d'Etude du Polymorphisme Humain) Utah) population.

### 3.3.2 Study population

The sampling of patients and controls is an important factor to consider in association studies since, in order to draw any general conclusions, we need to assume that our population of patients and controls are an unbiased sample of the underlying population [24]. Patients with an infrequent disease, such as MS, are often sampled at hospitals, and thus the controls should be collected within the same catchment area. The size of the study population has to be large enough to detect the anticipated effects, and now it is becoming increasingly clear that for MS the usual risk factors confer small increases in risk, which indicates the necessity to have large study populations.

All Swedish patients included in *Study I, II & IV* were recruited by neurologists at Karolinska University Hospitals, and all patients fulfilled the Poser [45] or McDonald criteria [3]. Blood samples are usually collected for clinical assessment of *HLA-DRB1* status and the patients are then asked whether they consent to participate in genetic studies. We have DNA samples from over 2000 MS patients, a number that gradually increases. In *Study I* 1084 MS patients with *HLA-A* and *HLA-DRB1* genotypes were included, excluding those who participated in the earlier *HLA-A* study [29]. For *Study II* we selected 973 Swedish and 484 Norwegian patients with sufficient HLA genotypes and clinical data. In *Study IV* samples from 1,077 MS patients were included, where the majority had been included in both *Study I* & *II*.

The blood samples from healthy controls were collected with the cooperation of different blood banks in the Stockholm area during two separate time periods: spring of 2001 and from late fall and spring of 2004/2005. Information regarding sex and year of birth was collected. In order to prevent double samples the blood donors during the second collection were asked if they had given blood at the blood banks used for the

first collection during the relevant time period. Blood donors were informed that samples would be used for genetic research at a Neurology department, and were not screened for MS. Additionally it has been argued that blood donors are healthier than the general population. These factors could possibly affect our possibility to detect true differences between patients and controls generalizable to the entire population. In *Study I* samples from 1,347 healthy controls were included, and in *Study IV* samples from 1,217 healthy controls were included.

### 3.3.3 DNA extraction

For all samples, blood was collected using EDTA tubes and frozen until DNA extraction. DNA can be extracted from almost every tissue. The usual procedure is to use blood as starting material because of the availability. At the Karolinska University Hospital Neurology clinics we have collected blood samples and extracted DNA since 1988, and consequently several techniques have been used in DNA extraction. When whole blood is used for DNA extraction, red blood cells are first lysed to separate them from nucleated cells, thereafter the white blood cells are lysed, proteins are removed and genomic DNA is precipitated. Different methodologies usually differ in how proteins are removed. Initially phenol chlorophorm extraction and methods where high salt concentrations are used to remove proteins were used at our clinics. Thereafter DNA was extracted using Qiagene kits (PAXgene™ blood DNA kit) where proteins are removed by incubation with a protease; this procedure generated DNA of better quality (less degenerated). Samples have also been extracted at the Karolinska Institute Biobank using Puregene kits (Qiagene) where a modified salting-out precipitation method is used.

### 3.3.4 Genotyping

Allele discrimination can be conducted using allele-specific hybridization, allele-specific primer extension, allele-specific oligonucleotide ligation or allele-specific enzymatic cleavage [46]. In the hybridization approach two allele specific probes are designed to hybridize to the target sequence only when they align perfectly [46,47]. Primer extension techniques are based on DNA template using either two sequence specific primers where only perfect matching primers produce amplified DNA, or one primer designed to anneal with the 3' end adjacent to the SNP site [46,47]. The nucleotide incorporated by DNA polymerase is determined by mass or fluorescence.

The ligation approach relies on the specificity of DNA ligase: two adjacent oligonucleotides are annealed to a DNA template and ligated only if there is a perfect match between both oligonucleotides and the template [46,47]. Alleles are then called based on the detection of ligation. The enzymatic cleavage approach utilize of the ability of different enzymes to cleave specific sequences or structures, and is used when the polymorphism affect such sites [47]. One example is the invasive cleavage approach which utilizes structure specific enzymes that cleave overlapping nucleotide sequences where the polymorphic site is at the point of overlap, and the overlapping structure is formed with the allele-specific probe [46].

The detection of specific alleles can be done by monitoring the light emitted by the products (fluorescence or chemiluminescence) or by measuring the mass of the products (by matrix assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS)) [46,47].

#### *3.3.4.1 Genotyping Study I & II*

Functional alleles of *HLA-A* and *HLA-DRB1* at a low (serological) resolution were identified using primer extension of sequence specific primers by Olerups SSP kits [48].

#### *3.3.4.2 Genotyping Study IV*

In *Study IV* SNP alleles were genotyped using one base primer extension technique at the Mutation Analysis Facility at Karolinska Institutet, and multiple SNPs were genotyped simultaneously. First, amplification probes amplified the region of interest, thereafter an allele specific extension reaction was conducted with the addition of dideoxy nucleotides. These primers were designed so that the difference in mass between all elongated primers in the multiplex assay was at least 20 Da. Alleles were thereafter detected using MALDI-TOF mass spectrometry. The genotype calls were manually checked by me and one employee at the core facility. Genotyping assays that failed at the first attempt were replaced with a tagging SNP when available, or else the assay was redesigned with new primers. Concordance analyses with the HapMap data as well as analysis of the parent-offspring-compatibility were performed, and showed concordance rates of above 99% for all analysed SNPs but rs570440 in *CD58* (98.3%) and rs8037783 in *HDGFRP3* (98.1%). All genotyping assays had a success rate above 89

% (mean 94 %), except three *HDGFRP3* SNPs that were excluded from the analysis (rs12441585: 63%, rs10162999: 60% and rs3816450: 29%).

### 3.3.5 Important concepts

Some statistical and epidemiological concepts will be discussed throughout my thesis, and are therefore presented and briefly discussed here.

- Null hypothesis: Here refers to that there is no association between the tested variant and the disease.
- Alternative hypothesis: Here refers to that the tested variant is associated with the disease.
- P-value: Probability of obtaining a result at least as extreme as the one observed, given that the null hypothesis is true.
- α: The significance level at which we are willing to reject the null hypothesis.
- Odds ratio (OR): Effect size. The odds of exposure among persons with the trait divided by the odds of exposure among individuals without the trait. This can be shown to be the same thing as the odds of trait among exposed divided by the odds of trait among unexposed.
- Power: Given that the alternative hypothesis is true, how likely is it that the test identifies the variant (given the effect size of the variant and the number of patients and controls).
- Type I error: Rejecting the null hypothesis when it is in fact true (False positive).
- Type II error: Not rejecting the null hypothesis when it is in fact false (False negative).
- Haplotype: The allelic variants along a chromosome in one individual.
- Hardy Weinberg equilibrium (HWE): If a population is in HWE, the genotype frequencies can be deduced from the allele frequencies so that at a bi-allelic locus: $f(AA) = f(A)^2$, $f(Aa) = 2*f(a)*f(A)$, and $f(aa) = f(a)^2$, where A and a denotes the two alleles. Deviance from HWE implies genetic drift, population stratification or selection [49].

#### *3.3.5.1 Linkage disequilibrium and Haplotypes*

Linkage disequilibrium (LD) is the non-random association of alleles at two or more loci [50]. Loci on the same chromosome usually have some degree of LD, which increases

with decreasing distance. The degree of LD is influenced by the rate of recombination between loci, but also by genetic drift, mutations, migration, population expansion and selection [51]. If alleles are in linkage equilibrium, all haplotype frequencies are equal to the product of all included allele frequencies. One basic measurement of LD, D, equals the difference of the observed (estimated) haplotype frequency and the product of the frequencies of the included alleles (see formula below). Usually in genetic case-control studies the phase (which alleles that are located on the same chromosome within an individual) is unknown and thus the haplotype frequency has to be estimated [52]. The D value is greatly influenced by allele frequency and therefore two different normalized measurements, based on D, are usually employed by geneticist: D' and $r^2$ [53]. Given two loci, A and B, with alleles $a_1$, $a_2$ and $b_1$, $b_2$:

$D = f(a_1b_1)-f(a_1)*f(b_1)$

$D' = D / D_{max}$

$r^2 = D^2 / f(a_1)*f(a_2)*f(b_1)*f(b_2)$

D' values can range between -1 and 1, $r^2$ values between 0 and 1, and $r^2 \leq |D'|$. $D_{max}$ is the maximum of D given the allele frequencies at the two loci. D' is always 1 if one of four possible haplotypes is missing, and indicates that the least common allele is always on a haplotype with *one* of the variants on the other loci, but does not necessarily mean that the allele at the other loci has any restriction in haplotype partner. A high D' can thus be interpreted as no recombination has occurred between the least common allele and the other loci. An $r^2$ of 1 signify that the two loci are *completely correlated* and one allele at the first locus only exists on one haplotype with one of the alleles from the other locus, and vice versa. The fact that alleles at loci close to each other often are correlated is utilized when performing genetic association studies, since it means that not all genetic variants need to be genotyped; highly correlated alleles capture the effects of each other.

LD can also be measured over several alleles at two loci using different methodologies. Cramer's V is a measurement where the statistic is based on a $\chi^2$ statistic (see below) and Kendall's tau-b measures the correlation between two rankings [54].

When a new mutation that change the susceptibility to a given disease occurs, this take place on a specific haplotype. The association between this haplotype and the mutated allele is only disrupted by new mutations and recombination. Therefore,

anonymous markers indicating this ancestral haplotype can be used to assess the effect of the disease-causing variant. It has been shown that DNA consist of blocks of higher LD where only a few haplotypes exist [55]; this can be due to the occurrence of recombination hotspots [56] or stochastic recombination [57,58]. Researchers have argued that since most of the genome exists within such haplotype blocks, the detection of alleles that tag these haplotypes can be used to assess all common variation [58,59]. Gabriel et al. [59] defined a haplotype block as a region where less than 5 % of the included pairs of SNPs showed strong evidence of recombination. Each of these regions are bounded by a pair of SNPs with 95 % confidence intervals (CI) for D' where the lower limit is above 0.7 and the upper limit is above 0.98. Inside these boundaries there are limits for the CI of D' between two, three, four or five markers, within specified regions of size depending of the investigated population. Wang et al. [58] defined haplotype blocks by using an extension of the four gamete test [60], where one investigate whether all possible haplotypes exist (if there has been recombination) and blocks are defined as regions of contiguous and ordered SNPs in which there is no evidence for recombination. A third method used within Haploview is the Solid spine of LD [61] which is an estimation algorithm based on a Poisson process model, penalized likelihoods, and cubic spline interpolation.

A common practise today is to perform a more conservative tagging, where one allele tags highly correlated alleles. A haplotype block definition is used to define regions within which association of haplotypes are tested in order to discover association of more rare variants that have arose on these more ancestral haplotypes. Other haplotypes than those restricted by high LD or low recombination could confer risk of disease over and above the risk conferred by any allele of a single SNP. Thus one might want to investigate all haplotypes within the genetic region, but due to restrains in statistical power and fear of false positives this is generally not conducted.

Since we usually do not know the phase of our alleles, algorithms to estimate haplotypes must be used. Normally the expectation-maximization (EM) algorithm developed by Dempster et al. [62], and introduced to haplotype estimation during the 1990ths [52], is employed. Here an initial guess of probable haplotype frequencies are supplied, usually the product of individual alleles, and thereafter an iterative process

takes places that finds the values for haplotype frequencies that optimises the probability of the observed data.

In *Study IV* all known (HapMap) SNPs within or close to the investigated genes were included in the tagging algorithm, and an $r^2 > 0.9$ defined a tagging allele. Haplotype frequencies were estimated by the EM algorithm in Haploview [63] and analysis was thereafter performed for haplotypes in regions defined by Gabriel et al. [59], or covering the entire gene.

#### *3.3.5.2 Genetic model*

Prior to investigating a certain marker for a complex disease, it is impossible to know how, if at all, this variant affects the trait. Carrying a certain variant could have an effect which is not influenced by the variant at the other chromosome: a *dominant* model. The opposite is called a *recessive* model, where both chromosomes must carry the variant in order for it to exert its effect. Logically, at a bi-allelic genetic marker these models are one and the same from the perspective of the different alleles. Often the effect depends on the number (zero, one or two) of specific alleles you carry: a *codominant* model. A more specific variant of this is the *multiplicative* model where the effect of being homozygote for the variant is the effect of carrying one allele to the power of two. Moreover, the heterozygote might confer effect when none of the homozygote does, an *overdominant* effect. A variant can also be X- or Y-linked when located on one of the sex chromosomes. A common procedure is to ignore the genetic models and perform the statistical tests on allele counts instead, which may be an biased approach if the case-control population deviates from HWE [64,65]. Alternatively, several genetic models could be tested in order to elucidate which one suits the investigated data better [66]. It could be argued that the correct genetic model would display greater statistical power and thus give the most significant p-value. In reality, the case-control cohort used is usually not big enough to make this decision, and the difference in significance levels might be minute.

In *Study I & II* we chose to investigate the multiplicative model, given that we know that HLA-DRB1*15 has a dose dependent effect [67,68] and that we thought that this was a reasonable assumption to make. The genotypic effects at these HLA loci would have been interesting to assess, but since 18 *HLA-A* alleles and 14 *HLA-DRB1* were identified

in our material, the number of genotypes to investigate was too large and the analysis would lack adequate power.

In *Study IV* we took a new approach, and chose to investigate five different models at once: recessive, dominant, codominant, overdominant and log additive (multiplicative) using the SNPassoc package [69] in R [70]. Thereafter the most significant model was reported and suggested to be the most accurate model to describe the SNPs effect on the trait, and was subsequently used in adjusted and interaction analyses. Undoubtedly this is a simplification of reality, and we do not have adequate power to draw firm conclusions regarding the best genetic model, but hopefully our results can persuade others in attempting the same thing and eventually the true genetic model will be known. In the case of the associated *CD58* SNPs, the multiplicative model produced the most significant results, but most other models also showed significant results. In contrary, *RPL5* showed most significant results using the overdominant model, but here both the codominant, and to a somewhat lesser extent the dominant model, also displayed significant p-values. Both the multiplicative and the recessive however lacked any sign of association. An association to the SNPs in *RPL5* and *CD58* could be detected using a test based on allele frequencies, which shows its robustness, but would not indicate the true genetic model if used solely.

#### *3.3.5.3 Interaction*

In genetic epidemiology, two different kind of interactions are usually discussed: biological and statistical interaction. The difference and meaning of the two can create confusion [71]. Biological interaction has been defined as when two risk factors together cause disease, or, using Rothman's analogy [24], are two pieces of the same pie. Rothman also showed algebraically that biological interaction results in deviation from additivity of the disease effects. Statistical interaction in the context of genetics can be defined as when two or more variants have an effect upon one another and can be tested in a regression model by adding an interaction terms into the model. As mentioned below, using logistic regression one assumes that the effects of variants are multiplicative, and thus biological interaction is assumed. There are exceptions, such as when there is a lack of biological interaction between two genetic risk factors then an interaction variable would usually be needed in the logistic regression model to represent this.

In neither of the included studies in this thesis was the biological interaction assessed specifically. I am however quite happy with assuming that there usually is biological interaction present between risk factors contributing to disease, especially since we assume that multiple risk factors contribute to MS susceptibility and that several sufficient causes exist.

In *Study I* we assessed the statistical interaction between alleles at *HLA-A* and *HLA-DRB1*, and in *Study IV* interactions between investigated SNPs and suggested risk modulators (female sex, *HLA-DRB1*15*, *HLA-A*02* and SNPs in *IL7R*, *IL2Rα*, *CLEC16A*, *CD226*, *SH2B3* and *KIF1B*) were assessed.

#### *3.3.5.4 Confounding*

Confounding is when a variable has an association with the variable you are interested in, *independent of the trait*, and additionally has an association to the investigated trait [72] (Figure 3).

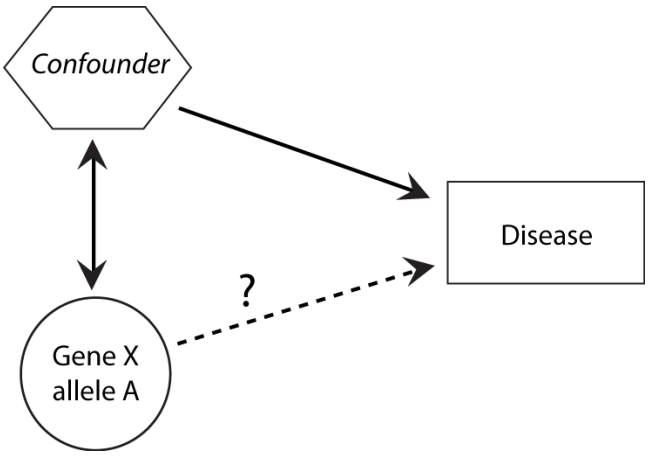


*Figure 3. Illustration of confounding. The confounder (a genetic or environmental factor) is associated to allele A of Gene X (which is the topic of investigation), and also associated to the disease. The association between Gene X and the disease is therefore biased by confounding, and might not reflect a true association.*

Thus the confounder obscures the true effect of the investigated variable to the trait. This can be avoided either by including both variables in a logistic regression (see below), by selecting patients and controls that have equal distributions of the confounder or by adjusting the effect measures using the Mantel-Haenszel methodology [24]. When using the Mantel-Haenszel methodology an estimate of the

common effect of the variable across the confounder strata is calculated using a weighted mean, but should only be used when the effect sizes in the different strata are homogeneous [72]. If the effect size differs in the different strata one is facing interaction, and other routes of analysis should be taken (see above).

A confounder in a genetic association setting can be another allelic variant in LD with the variant you are interested in, but could also be an environmental factor associated to the genetic variant you are investigating. A more far-fetched situation would be that sex would be associated both to the trait and your investigated genetic variant. This could occur if the genetic variant is present on either sex chromosome, but also if the variant worsen the survival in one sex but not the other, and hence would become more common in the second group.

The existence of confounders is hard to argue against, but they are usually unknown, although all previously known risk factors for a trait could possibly be confounders.
Matching of patients and controls for some possible confounders has been intensely argued especially for avoiding population stratification [73]. Recently however, the collective impression is that population stratification is not a major problem in well powered studies [32] and a gain in efficiency when stratifying for ethnic background is only apparent if there is a strong confounding effect [74].

Showing an effect of a variant while adjusting for known risk factors or possible confounders, and statistical interaction with these, implies that the variable in question has an effect independent of those possible confounders. Usually if a gene is located close to a known genetic risk factor there are reasons to suspect that the risk conferred by the known risk factors might influence the attained results of the investigated variant. Therefore it is important to try to assess possible confounding effects, and interactions, in each genetic study. In *Study I & II* we adjusted for the possible confounding effect from *HLA-DRB1* alleles on *HLA-A* alleles and vice versa. In *Study IV* we assessed the independence of each SNP from other suggested risk factors.

### 3.3.6 Statistical tests

*3.3.6.1 $\chi^2$, Fisher's exact and Rank tests.*

In studies investigating genetic association, the question is simply: "Does the distribution of the studied variant differ between individuals with and without the trait?" If is does, the variant is said to be associated with the trait. As discussed above, one might want to analyse the genotype distribution, the carriage of an allele or the heterozygote effect. If your data easily can be tabulated into a contingency table, i.e. the carriage of allele A among the affected and unaffected (Table 1), both a $\chi^2$ and Fisher's exact test is applicable, where the null hypothesis is that the two variables are independent of each other.

| | A + | A - | Total |
|---|---|---|---|
| Patients | $O_{11}$ | $O_{12}$ | $R_1$ |
| Controls | $O_{21}$ | $O_{22}$ | $R_2$ |
| Total | $C_1$ | $C_2$ | n |
| | $C_1/n$ | $C_2/n$ | |

*Table 1. Schematic 2x2 table.*

*$OR = (O_{12} * O_{21}) / (O_{11} * O_{22})$*

*Expected frequency of carriers of allele A under the null hypothesis of no association between A and disease is $C_1/n$.*

Fisher's exact test is a nonparametric exact test where the margins are fixed. As test statistic the smallest value in the 2x2 table is used, and under the null hypothesis the probability of the observed results has a hypergeometric distribution.

The $\chi^2$ test is based on the assumption that the test statistic has a $\chi^2$ distribution under the null hypothesis. Using large samples, this assumption is usually met, but when any count is low the Fisher's exact test should be employed instead [72]. In the $\chi^2$ test the statistic is based on the difference between the observed allele frequencies and the values of $C_i/n$ which would be equal under the null hypothesis, assuming HWE.

Rank tests are non-parametric test that do not assume any distribution of the investigated variables except that all investigated variables have the same (possibly dislocated) distribution. Here the combined values in all groups are ranked, and the positions of each group values are evaluated. The Kruskal–Wallis rank sum test is an extension to the Mann-Whitney U test when investigating more than two variables. Here the statistic is based on the difference between the mean rank in each group and the total average rank [75]. Several additional highly similar methods, such as Spearman's rank correlation coefficient, exist.

In *Study I* the Kruskal-Wallis rank sum test was used to assess whether HLA alleles influenced severity among MS patients as assessed by MSSS. The effect of HLA alleles on MS susceptibility was investigated primarily using logistic regression (se below), but was additionally compared with results produced using Fisher's exact test.

*3.3.6.2 Regression*

Another alternative for finding the association between one, or *several,* variables and an outcome is regression analysis. Linear regression investigates whether a continuous outcome is dependent the variables in question according to:

$Y = \beta_0 + \beta_1 * X_1 + \ldots + \beta_n * X_n + \varepsilon$

Where $Y$ is the dependent outcome, and $X_i$ is the variables we are investigating, $\varepsilon$ is a normally distributed error term, $\beta_0$ is the intercept, or baseline, and $\beta_i$ is a parameter measuring the effect of $X_i$ on Y. Given the data all $\beta_i$ are estimated using the least squares criterion and thereafter subjected to t-tests [72]. In *Study II* we used linear regression to test whether alleles at *HLA-A* and *HLA-DRB1* had an effect on age at onset of MS, and the $\beta_i$ would in this case represent the difference in age among those carrying the $X_i$ allele as compared to those carrying the alleles utilized as baseline.

Many techniques have developed from simple linear regression, and involve different transformations and link functions. The Cochran-Armitage Trend test has been used for assessing significance of genotypic data and is an appropriate test when the independent variable is ordinal. Here the *probability* of the outcome (e.g. disease) is assessed as linearly dependent on genotype. Logistic regression can also be used when the outcome variable is dichotomous, here the dependent variable is based on the probability of disease but has been transformed with a link function: *ln(P/(1-P)) = ln(odds)*, usually denoted logit(disease). Here all $\beta_i$ are estimated using least squares or maximum likelihood. The significance of the model can then be assessed using a likelihood ratio test (among others), and the significance of one variable (estimator) in the model can be assessed using the Wald statistic. The Cox Proportional Hazards model, is a sort of a regression model on survival data, investigating the effect of several variables on time (dependent variable) to a specific event [76].

The advantage of regression models is that several independent variables can be determined simultaneously, and thus the independent effect on one variable from another can be assessed, and all estimated effects are adjusted for other included variables. One allele at each loci will by necessity be included in the intercept of the model, and estimated effects would be relative to this baseline. This is usually not a concern when analysing SNP alleles, but might cause confusion when multiallelic markers are assessed. Therefore caution should be employed when selecting intercept, especially when dealing with a multivariable locus, and one line of action is to select a common allele with equal frequency among individuals with and without the disease.

In *Study I* we used logistic regression modelling to show that the *HLA-A* locus has an effect on MS susceptibility *independently* of *HLA-DRB1*, even though some LD exists between these loci. Initially a model where all *HLA-DRB1* alleles were used to predict probability of MS was evaluated. Thereafter a model including both *HLA-A* and *HLA-DRB1* allels was also evaluated, and thereafter compared using a likelihood ratio test to the initial model. Additionally, models including interaction variables between alleles at the two loci were investigated as well. Thereby, we were able to conclude that the *HLA-A* effect is neither caused by a confounding effect nor an interaction effect with *HLA-DRB1* in our material. In *Study I* we additionally compared our results for each allele using multivariate logistic regression with those produced by Fisher's exact test and Cochran-Armitage Trend tests. As expected, for some alleles the different analyses yielded conflicting results, e.g. the HLA-A*03 allele is associated using both Fisher's exact and Cochran-Armitage Trend tests, but showed no sign of association when effects were adjusted for other included alleles (using logistic regression).

In *Study II* logistic regression was used to investigate the influence of *HLA-A* and –*DRB1* alleles on disease course, and linear regression was used to investigate the influence of HLA alleles on age at onset.

In *Study IV* we also assessed association to MS susceptibility and course (bout onset or PPMS) using logistic regression, using five genetic models (recessive, dominant, codominant, overdominant and multiplicative) for each SNP. In this analysis we were able to confirm a role for *CD58* and the *RPL5* region in MS susceptibility. Moreover, using the most significant genetic model for each SNP in *CD58*, *HDGFRP3* and the

*RPL5* region we also investigated how previously known risk factors interact and affects the association to MS susceptibility. Survival analysis on time to EDSS six using Cox Proportional Hazard model was also performed, but showed no influence of *RPL5*, *CD58* or *HDGFRP3* on severity.

### 3.3.7 True effects

When do we believe that a genetic association is true? I argue that the only convincing evidence is replication. For many years several association studies in many diseases have been published, but almost none have been confirmed in a second publication. This is most probably due to type I errors, but likely also type II errors in the small follow up studies. Using a medium sized case-control material of 1000 patients and 1000 controls many of the now confirmed risk alleles for MS would often only show significance levels of $10^{-2}$. Those values would normally not survive multiple testing corrections if a reasonable number of markers were included simultaneously. Thus it is clear that strict correction for multiple testing will produce type II errors. In my mind, type II errors are worse that type I errors, since a type I error might be discovered in the coming follow up studies, while a initial possible false negative finding will escape validation. It is however clear that caution *always* should be employed when interpreting novel findings, especially when associations are found in small populations or in stratified subpopulations. It can always be argued that effects seen are due to confounders, known or unknown, and thus the independence from previous risk factors might be useful to investigate. Once all our patients and controls have been completely sequenced, we might be able to prove that genetic effects are independent of confounding effects due to LD, but we still cannot exclude confounding by environmental factors. When convincing data in favour of a genetic effect i presented, the biological cause of association should be explained and proven. This daunting step has never been completed in any complex disease as far as I know, and could possibly require more complete knowledge on the complex molecular interplays occurring in an individual.

## 3.4 SUMMARY OF RESULTS OF STUDY I, II AND IV.

Four genes have been investigated throughout this thesis, and we have showed associations with MS susceptibility for *HLA-A*, *CD58* and the *RPL5* region.

In *Study I* the *HLA-A* association was shown to be independent of the long known *HLA-DRB1* association. At the *HLA-DRB1* locus the DRB1*15 allele conferred the largest effect: an OR of 2.3 per allele, p-value of $8x10^{-9}$. All other DRB1 alleles showed neutral or protective effects where DRB1*01 and DRB1*X (mainly attributable to DRB1*07) were significantly protective. At the *HLA-A* locus only the HLA-A*02 allele showed significant association to MS susceptibility, with an OR of 0.65 and p-value of $1x10^{-4}$. Moreover, we also evaluated interaction terms between *HLA-A* and *HLA-DRB1* alleles, between A*02 and DRB1*15 in particular, that did not reveal any significant interaction. Thus, the *HLA-A* effect is neither caused by a confounding effect nor an interaction effect with *HLA-DRB1*.

In *Study II* we could not detect any effect of *HLA-A* on disease course, age at onset or severity (as measured by MSSS), whereas an association between HLA-DRB1*15 and lower age at onset was confirmed.

When investigating *CD58*, *HDGFRP3* and *RPL5* in *Study IV* we were able to control for possible confounding effects by the newly discovered risk factors *IL7R*, *IL2Rα*, *CLEC16A*, *CD226*, *SH2B3* and *KIF1B* as well as female sex, *HLA-DRB1*15* and *HLA-A*02*. The associations to *CD58* and the *RPL5* region were confirmed in our Swedish case-control material, and were not secondary to any other known risk factors. *RPL5* showed a heterozygote effect on MS susceptibility, whereas *CD58* had a multiplicative effect. Moreover, we were able to show suggestive interactions between SNPs in *RLP5* and sex, which, through a stratified analysis, indicated that *RPL5* contribute to MS susceptibility among men, but not women. The *RPL5* region contains three genes that have showed association to MS susceptibility: *EVI5* (ecotropic viral integration site 5), *FAM96A* (family with sequence similarity 96, member A) and *RPL5*, and the relationship between these genes in MS require more investigation since they are in LD with each other. Additionally, we showed some suggestive associations between SNPs in *CD58* and *HDGFRP3* to the primary progressive course of MS. Some associations, such as an interaction between *RPL5* variants and sex, and associations between *HDGFRP3* and *CD58* variants and disease course are more suggestive, and warrant further investigation. Survival analysis on time to EDSS 6 using Cox Proportional Hazard model was also performed, but showed no influence of *RPL5*, *CD58* or *HDGFRP3* variants on severity.

## 3.5 DISCUSSION ON MS GENETICS

During the 1970[ies] several HLA class I alleles were reported as associated, those with increased frequency among MS: HLA-A*03, A*07, A*10 and B*07 and those with decreased frequency among MS patients e.g. HLA-A*02 and A*12 [2,26,77,78]. Later, a strong association between the *HLA-DRB1* allele DRB1*15 and MS was described and the initial class I associations were regarded as secondary [27]. In a later meta-analysis by Jersild [2] HLA-B*07 was associated to HLA-DRB1*15, indicating that a B*07 association could be secondary to the stronger DRB1*15 effect. No specific association was found between DRB1*15 and A*03, even though A*03 and B*07 showed association. Thus, the associations of HLA class I entities were never exhaustively investigated, but simply regarded as secondary to the association between B*07 and DRB1*15.

MS is associated with a combination, haplotype, of HLA class II alleles: DRB1*1501, DRB5*0101, DQA1*0102, DQB1*0602 [28]. The DRB1*15 allele is present in about 55-60 % of MS patients and 30 % of healthy individuals (our data), and is associated to MS in practically all populations studied. Additional studies have proposed roles for other alleles at *HLA-DRB1*, or combination of certain alleles in MS susceptibility [79-81].

Due to the early proposed roles of HLA class I alleles in MS susceptibility, and reports on the effect of class I molecules in EAE, these genes were investigated within our group by Fogdell-Hahn et al. [29], and later Harbo et al. [30]. The effects of *HLA-A* alleles were investigated through stratified analysis according to carriage of HLA-DRB1*15, and showed somewhat conflicting results on whether A*03 or A*02 possessed effects in all subgroups. To elucidate the role of *HLA-A* entities we used a large independent study population in *Study I* and determined possible associations using logistic regression, thus enabling *HLA-DRB1* adjusted statistics to be assessed for *HLA-A* alleles. Thereby we confirmed the strong protective effect of HLA-A*02, while the crude HLA-A*03 association did not reach significance when adjusting for other alleles. Thus there is clearly a genetic factor influencing risk of MS in the HLA class I region, but there are conflicting arguments regarding the responsible gene: Yeo et al. later published a paper using stratified analysis where HLA-C*05 is suggested to be associated independent of other class I and II alleles [82]. The final strata in this analysis where the

*HLA-C* association was observed was however small, 226 MS patients, and it was unclear in the report whether this allele was associated in the entire population or only in the final strata. It remains to be elucidated whether multiple loci within the HLA class I region affects MS susceptibility. In order to perform an exhaustive investigation of effects in the class I region, more genes need to be assessed, tentatively *HLA-B*, but many other functional candidates exist. Most MS researchers however agree that there is a true MS locus in, or close to, the HLA class I region.

At the *HLA-DRB1* locus two additional variables showed significant association in *Study I*: DRB1*01 and DRB1*X (consisting of rare alleles) mainly attributable to DRB1*07. Even though we are able to adjust for effect at other loci, the reciprocal effect of associated alleles is not avoided using logistic regression. By sequentially excluding individuals carrying the most significant allele at the *HLA-DRB1* locus we nevertheless saw that both DRB1*01 and DRB1*07 contribute to the modulation of risk for developing MS (data not shown). Thereby we also confirm that several alleles at the *HLA-DRB1* locus affect disease susceptibility. In contrary to some studies [79,80,81 ] we do however see an independent effect of DRB1*01, whereas others suggested that DRB1*01 was protective only in the presence of DRB1*15. Independent effect of DRB1*17 and DRB1*14 have been suggested [79,80,81 ] but were not confirmed in *Study I*.

In *Study I* we illustrated the biological interactions of being a carrier of different genotypes at *HLA-A* and *HLA-DRB1*, where the most susceptible genotype combination conferred an OR of 22, and the most protective an OR close to 0.4.

The HLA molecules are responsible for presenting the majority of all present peptides within the individual, this is accomplished by two routes. Firstly, via the existence of multiple genes of both the class I and class II type, arising from earlier occasions of gene duplication (reviewed in [83]). Additionally, the HLA genes are highly polymorphic, and one specific HLA molecule binds a range of different peptides, sharing only a few conserved residues [84]. HLA class I and II molecules are responsible for antigen presentation to T-cells, but possess different roles. HLA class I molecules are present on all nucleated cells and mainly present peptides produced within the cell to CD8+ T-cells, which are programmed to kill those cells they specifically recognize. HLA class II molecules present peptides to CD4+ T-cells, whose role is to activate other cells

of the immune system. Thus, HLA class II molecules are present on those cells participating in eliciting immune responses, such as dendritic cells, B-cells and macrophages. Unlike class I molecules, class II molecules bind peptides originating from phagocytosed proteins present in acidified endocytic vesicles [84].

As with any genetic association, a true associated variant could confound our findings, by being in LD with the detected variants. This possibility will not be disproven prior to the resequencing of large numbers of MS patients and controls, and will therefore be mostly disregarded during the remainder of this discussion. The possible causal role of variants of both HLA class I and II are easily argued for due to their fundamental role within the immune system, but so far the knowledge on mechanisms for how the specified variants confer altered risk of MS is limited. MS associated molecules could facilitate the presentation of encephalitogenic peptides to T-cells, or compromise the negative selection in the thymus. Otherwise, the current HLA molecules might have different affinity for peptides resembling CNS antigens (molecular mimicry).

Interestingly, the HLA-A*02 allele has the ability to present ligands independent of the TAP (Transporter, ATP-binding cassette) complex that usually transport peptides into the ER [85], and instead bind signal sequence-derived peptides which have been released into the ER by the signal peptidase complex [86]. This indicates that the HLA-A*02 allele possesses quite unique functions, and might thus partly explain the association with this specific allele. HLA-A*02 is able to bind peptides created under stressful/deviant circumstances, as when various viruses [87] try to evade the immune surveillance.

In an elegant study by Friese et al. [88] mice transgenic for either human HLA*03 or double transgenic for A*03 and A*02, and/or transgenic for a T-cell receptor (TCR) specific for the myelin proteolipid protein (PLP) 45-53 (2DI-TCR), presented by HLA-A*03 [89] were investigated. Double transgenic mice (A*03 and 2DI-TCR) showed symptoms of disease after immunization with PLP 45-53, albeit not single transgenic mice. Double transgenic mice showed early infiltration of mostly CD8+ T-cells binding the A*03-PLP complex, but about 15 % were CD4+ T-cells. Later in disease CD4+ T-cells dominated in the CNS, and disease progression was shown to be dependent on functional CD4+ T-cells. Interactions between class I genes were investigated, and the

addition of an HLA-A*02 transgene to A*03 and 2DI-TCR double-transgenic mice completely prevented disease. The spleens of the triple transgenic mice showed a 90 % reduction of CD8+ T-cells able to bind the A*03-PLP complex. These results indicate that expression of HLA-A*02 results in negative selection of 2DI-TCR thymocytes that express a higher level of this TCR receptor. It remains to be elucidated whether this occurs in humans as well, and whether HLA-A*02 contributes to negative selection and reduced level of other autoimmune clones as well, but that might be hypothesised.

In *Study II* we did not detect any association of alleles at *HLA-A* to age at onset, disease course or severity (as measured by MSSS). We did however validate an association between HLA-DRB1*15 and younger age at onset, as previously suggested [90-93], although our material is partly overlapping with the Masterman et al. and Celius et al. study. Additionally, we found an association between DRB1*04 and PPMS prior to correction for multiple testing, as also suggested previously [94-96]. The PPMS groups are however always small, making these results more difficult to argue for.

The apparent lack of correlation between possessing an effect on MS susceptibility and possessing a role in MS severity may seem counter intuitive. A variant influencing whether disease develops in an individual could be expected also to take part in the expressitivity above the limit of detection, and would therefore also be evident when assessing severity. One must bear in mind however that case-control cohorts have just recently grown to the size where genetic associations can be detected at the effect size that is common in complex diseases. Therefore, performing statistical analysis on half this population (the patients), for phenotypes that may be even more difficult to pinpoint than the actual MS diagnosis, has a limited power. I believe that when larger patient cohorts are available, and we as researchers have begun to use the proper tools for assessing severity, MS susceptibility genes will in many occasions also be found to influence disease severity.

A few years ago the first non HLA gene in MS susceptibility, *IL7R* [34,36], was discovered within our group and subsequently confirmed in other populations [33,35]. Through the IMSGC screen [33] many new candidates were brought to life, and *IL2Rα* [38,97], *CLEC16A* [38,43,98], *CD58* [38,42](*Study IV*) and the *RPL5* region [38,39](*Study IV*) have now been confirmed as genes conferring MS susceptibility. Furthermore the

IMSGC investigated SNPs that had shown associations to type I diabetes, and thereby showed associations to *CD226* and *SH2B3* and confirmed the association to *CLEC16A* with a second SNP [43]. Recently, a study initiated in a Dutch isolate and further investigated in case-control materials from Holland, Canada and Sweden suggested a association between *KIF1B* and MS [41]. Within our group we attempt to genotype all discovered genetic risk factors for MS in our cohort, and due to this I was able to investigate our candidate genes in relation to other factors thought to affect MS susceptibility. Thereby, twelve factors were mutually investigated: sex, *HLA-DRB1, HLA-A, IL7R, IL2Rα, CLEC16A, CD58, RPL5, FAM69A, CD226*, *SH2B3* and *KIF1B*, as well as all two-way interactions.

In *Study IV* we tried to map the effects within *CD58*, *HDGFRP3* and *RPL5*, and also included two SNPs in *FAM69A* since its nucleotide sequences are overlapping with *RPL5*, and both of them were found to be differently expressed in *Study III*. The association to *RPL5* was confirmed, but several SNPs showed association, both in *RPL5* and *FAM69A*, and these variants were in high LD (D' above 0.9, $r^2$ between *RPL5* variants and *FAM69A* variants at 0.26) in our data. Furthermore, we did not include the *EVI5* variants that have shown association [39] and these could also confound our finding, even though they are contained within separate LD blocks according to HapMap. The risk seemed to be conferred by the heterozygote, and moreover we detected a suggestive interaction between *RPL5* and sex implying that *RPL5* may be a risk factor among men only. *RPL5* encodes a ribosomal protein in the 60S subunit and thus takes part in the translation within cells. It could be so that different variants of these proteins can translate different proteins more or less effective, and thereby affect MS susceptibility.

CD58 is a co-stimulatory molecule that acts as receptor for the surface CD2 antigen present of T and NK-cells. Therefore different variants of this molecule might influence the affinity and affect the following immune response. CD58 is expressed on endothelium in brain micro vessels in the blood brain barrier (BBB) and therefore impacts the ability of T-cells to cross the BBB (reviewed in [99]).

Many of the genes contributing to MS susceptibility are immune related: *HLA-A*, *HLA-DRB1* and *CD58* as mentioned above, IL7Rα is crucial for proliferation and survival of T

and B lymphocytes [100], IL2Rα is a T-cell growth factor receptor involved in the suppression of autoimmune disease, *CD226* encodes a membrane molecule involved in the adhesion and co-stimulation of T-cells and *SH2B3* encodes an adaptor protein mediating the interaction between the TCR and intracellular signalling pathways. Less is known about *CLEC16A* but it is expressed on B-cells, NK-cells and dendritic cells. Different variants of these genes can thus be hypothesised to affecting MS susceptibility by changing the affinity or intensity of reactions conferred. Other associated genes are not immune related like *RPL5* as mentions earlier, or *KIF1B* encoding a protein involved in the transport of mitochondria along microtubules.

Many additional genetic risk factors will probably be found within the next couple of years, and those we know today might be redefined. Now might be the time to start indulging ourselves with functional characterization of these genetic variants.

# 4 GENE EXPRESSION PROFILING

Gene expression profiling is a technique for simultaneous describing mRNA levels of multiple genes, thus trying to deduce which processes the investigated cells are involved in. Gene expression profiling captures a snapshot of the activities within sampled cells, at the mRNA level.

Whereas the genetic code within an individual stays rather constant during life, the expression of genes is known to vary between tissues and throughout life. Therefore the investigation of gene expression might be more difficult. At the same time, the variation between individuals at a certain gene is larger at the mRNA level than at the DNA level which implies that gene expression might explain the difference in phenotype to a larger extent than DNA. The changes in gene expression can however also often be the result of disease, direct or indirect, adding another level of uncertainty.

## 4.1 AIM

The aim in *Study III* was to investigate possible differential expression of genes both centrally and peripherally in MS patients compared to controls, in order to formulate hypotheses about ongoing processes.

## 4.2 DIFFERENT TECHNOLOGIES

Several technologies exist for detecting the levels of mRNA of multiple transcripts simultaneously, and generally include either a sequencing approach [101] or a large number of nucleotide probes attached to a surface; the latter approach is used within this thesis and henceforth discussed. The sample of interest is labelled and then hybridized to the probes and subsequently detected. Different methods differ in the length of the probe on the surface, how the probes are synthesized, how many samples are hybridized on one array etc. Measuring the quantity of label in each location gives an intensity value that should be correlated to the quantity of the corresponding transcript in the sample. There are two major routes of labelling in use today. One can hybridize two samples on each array labelled by a red and a green fluorescence, and then measure the ratio of colour emitted. Alternatively, one sample is labelled with a fluorescence label and then hybridized to a single array, and the

intensity of each location is measured. In our gene expression study (*Study III*) we utilized arrays manufactured by Affymetrix (Santa Clara, CA, USA) where a single sample is hybridized to each array.

## 4.3 AFFYMETRIX GENE CHIPS

The oligonucleotides on the Affymetrix arrays are designed *in silico* and synthesised on the array surface using a photolithographic procedure, contributing to a low batch-to-batch variability. Probes are either designed against the transcripts 3′ end [102], or against each known exon in a gene [103]. The latter approach consequently has the potential to detect and discriminate between different transcription variants, known or unknown. Additionally, tiling arrays are available in order to discover novel transcripts [104]. *Study III* was conducted using microarrays with probes designed against the 3′ end of the transcripts; hence further introduction will only discuss matters applying to that approach. In order to detect and correct for unspecific binding to probes, there is one mismatch (MM) probe for each perfect match (PM) probe on many Affymetrix array, these differ in one base on the 13$^{th}$ position. Each oligonucleotide on the array is 25 nucleotides long, which would generate a rather low specificity if used alone. Therefore a single transcript is detected by, at least, one *set* of probes, with 11 pairs of probes in each set. We used GeneChip® Human Genome U133 Plus 2.0 arrays in *Study III* which contain more that 54,000 probe sets and has a feature size of 11 µm. At the time of probe design of the U133 arrays, probes were designed to bind all known transcripts found in UniGene with information from a draft assembly of the human genome from the University of California, Santa Cruz (April 2001). A draft assembly of the human genome from NCBI (Nov 2003) was used to design the additional probe sets on the Plus 2.0 arrays. In total, 38,572 UniGene sequences, 2,669,196 bdEST sequences, 49,135 GeneBank sequences and 13,696 RefSeq sequences were used to design the probe sets on the utilized arrays [105]. These arrays detect more than 47,400 transcripts of more than 38,500 genes [102], figures that vary due to that the Affymetrix annotation is built on UniGene clustering which are regularly updated, and never manually curated (personal communication, NCBI Helpdesk).

### 4.3.1 Study population

In analogy to the discussion regarding choice of patients and controls in genetic studies, we want our study population to reflect the entire population of individuals

with disease and the general population. Gene expression profiling studies are however usually smaller than genetic studies, due to financial restrictions. One might assume that gene expression fluctuate more, and has a larger variation between individuals than genetic status. The variability within one individual, between tissues and within tissues is huge [106].

For *Study III* we included 12 individuals with MS that were sampled during a disease bout, 14 MS patients sampled during remission and 18 controls with other neurological diseases (OND) where all controls had a non-inflammatory disease. Samples from individuals in disease bout were sparse, and by chance a majority of the samples with good enough sample quality and clinical characteristics were from males. Since our most profound aim was to investigate differences between MS patients and controls we matched these groups according to sex. All samples were collected prospectively as a part of a larger undertaking, during scheduled visits at the Karolinska University Hospital's neurology clinics. All study participants gave their informed consent, and were not treated with any immunomodulatory treatments.

### 4.3.2 What tissue is relevant to investigate?

Given a specific disease, one might ponder on which tissue that is most relevant to investigate. In *Study III* our aim was to investigate the processes that occur within an MS patient, and one obvious choice of tissue might be lesions from within the CNS. Removal of tissues from the brain of living individuals is not ethically defendable, and the use of post mortem samples would usually imply late stage disease, high age or unusual circumstances in addition to the changes in RNA occurring after death, all of which could influence the microarray investigation [107]. In our case, we have a biobank of CSF samples from newly diagnosed patients, and patients with suspected or confirmed neurological diseases of other kinds. CSF has long been used as a surrogate of tissues from the brain and brainstem, and it does not seem unlikely to assume that changes within the CNS would affect the CSF as well. The CSF does however contain a low amount of cells, and large quantities were usually needed in order to perform a gene expression profiling investigation. Improved purification and amplification procedures have decreased the demands on amounts of RNA, and we showed that from as little as 0.68 ng of RNA we were able to produce an adequate

amount (>15µg) of labelled cRNA via a two-cycle labelling protocol, which enabled us to proceed with *Study III* as planned.

CSF is produced at a number of sites within the brain, in particular in the choroid plexus, and is a watery solution that serves as a transport medium. Lymphocytes perform immune surveillance in the CSF, and about 80-90 % of the present cells are T-cells [108]. In the CSF of normal healthy individuals the CD4+ T-cells dominate, whereas mostly CD8+ T-cells are present in brain tissue (reviewed in [99]). The cellular composition of CSF is not a simple reflection of cells in peripheral blood [108], thus migration into CSF is controlled. Active lymphocytes migrate to and from the CSF from blood trough the blood CSF barrier, and to/from the CNS through the CNS CSF barrier, and a recent report suggest that lymphocyte entry occurs via the choroid plexus [109]. Lymphocytes within the CSF and CNS of MS patients are believed to have migrated from blood, and selective migration, selective accumulation and/or clonal expansion of cells creates antigen and clonally-restricted populations [110-113]. The composition of cell populations in CSF has been shown to be different between MS patients and controls (with non-inflammatory neurological diseases); MS patients have larger proportions of B-cells and plasma cells, and lower proportions of monocytes, natural killer (NK)-cells and NK-like T-cells [108]. No correlation between cell populations in CSF and PBL could be seen in that study, and no statistical difference between the two groups was detected in PBL. In addition, the distribution of cell populations in CSF of MS patients remained stable through time and exacerbations [108].

The sampling of CSF through lumbar puncture is an unpleasant event for the subject. Blood on the other hand is easily accessible, and therefore an easy choice when performing any biological study on a supposedly autoimmune disease. One of our aims was to investigate whether any biomarkers could be identified, genes with expression levels that clearly separate MS patients from controls, which would also be more useful to discover in cells from blood than from CSF cells. Moreover, there have been claims that blood mimics the processes in the CNS in MS [114]. We chose to perform global gene expression profiling in two tissues from each individual simultaneously in *Study III*, both CSF and PBL. Gene expression profiling has been conducted earlier in MS using cells from blood [115-122] and solid brain tissue [123-127].

It is clear that differences in gene expression in any experiment, using any technique, in any tissue examined can be due to different cell compositions in the samples examined. The investigation of gene expression does however give another resolution of processes occurring than investigating differences in cell populations. My personal opinion is that multiple studies in different tissues, sorted cells from tissues and cell lines, using different techniques and inquiring different levels of processes will be necessary in order to elucidate the processes occurring in MS.

### 4.3.3 Sample preparation and hybridization

Sample handling, RNA extraction and storage can affect RNA quality and thus the microarray results. In *Study III* we assessed RNA quality prior to microarray procedure using 28s/18s rRNA ratio and RNA integrity number (RIN) [128] as measured by a 2100 bioanalyzer (Agilent, USA). In a study by Thompson et al. [107] RIN values below seven were considered low quality, and sensitivity in microarray studies was markedly decreased at this level.

For *Study III* both CSF and PBL samples were collected at the same scheduled visit at the Karolinska University Hospitals Neurology clinics during 2002 until 2006. CSF samples were immediately centrifuged, and the pellet was recovered and stored at -70°C until use. The blood peripheral lymphocytes were separated, pelleted and frozen on dry ice and stored at -70°C until use. Total RNA was extracted using PicoPure™ RNA isolation kit (Arcturus, USA) according to manufacturers instructions, and with DNase treatment according to supplier's instructions (Qiagen RNase free DNase set, Hilden, Germany). We demanded a high RNA quality as measured by a18s/28s ratio above 1.3 and a RIN value above 7.7.

Since cell numbers are sparse in CSF, we investigated how low quantities of total RNA that would be sufficient to produce an adequate amount of cRNA for hybridization (>15µg) when using the Affymetrix two-cycle labeling protocol (Affymetrix, Santa Clara, CA, USA). Six low quantity samples ranging from 0.14 to 3.18 ng of RNA were tested, and only the lowest amount failed. Thus, we concluded from as little as 0.68 ng of RNA enough cRNA can be produced.

We chose to treat all included samples equally, so even though most PBL samples had plenty of RNA they all underwent the two-cycle labeling procedure, and all but two samples produced adequate amounts of labelled cRNA. All samples were hybridized to Human Genome U133 Plus 2.0 arrays, and were thereafter visually inspected in true colour and using different scaling, and two arrays with regional biases were excluded (one MS CSF, one control PBL).

### 4.4 PRE-PROCESSING

Pre-processing involves background correction, summarization of probe sets and normalization. The aim is to analyse true biological differences, and not variance introduced by sample preparation, array manufacturing or processing (labeling, hybridization, and scanning). There is known correlation between variance and mean, where low intensities lead to large variance [129,130]

Several methods exist for all pre-processing steps, and I will focus on a few of those that apply to Affymetrix arrays. The most influential step in pre-processing is background correction [131], dealing with unspecific binding; this correction can be either probe specific or not. As mentioned earlier Affymetrix probes usually come in pairs of perfect matches and mismatches. The original intention was that background correction could be accomplished by subtracting the MM values from the PM values (this is performed in the MAS 5 algorithm). In reality up to one third of probe pairs on a given array have MM values that are higher than the PM values. Additionally values of MM grow with PM values, indicating that MM probes detect the same transcripts as the PM probes [129,132].

Strong probe effects exist for Affymetrix arrays where a probe set usually shows a distinctive profile across different arrays, suggested to be due to different affinity or position dependent base effects [133]. Although different treatments or disease stages might increase the variance in probe set profiles [134]. These effects demand an adjusted method for summarizing the intensity values from the 11 probe pairs, available methods include: robust average, linear mixed models, multiplicative models fitted to PM or PM-MM, trimmed mean and others [129,132 ,135].

Lastly, normalization is intended to balance the individual intensity values, since small differences in e.g. total RNA quantity or hybridization periods might affect overall intensities. Affymetrix original scheme (used in MAS 5) was to scale each array so that the mean intensity for all arrays is the same; this gives a scaling value that is used in later quality control (QC) steps.

Through literature studies [129,131,132,136,137] at the start of *Study III* we decided that the most favoured methods for conducting pre-processing was robust multi-array average (RMA) [133] and CG-RMA [131]. These methods both use a robust linear model to summarize probe sets, but perform different background corrections. RMA uses a global background correction, while GC-RMA includes information about GC content in probes. None of them use the MM probe data, and both work in log2 scale where PM values grow roughly linearly with respect to concentrations [131]. Normalization in both these methods assume that the majority of genes are not differently expressed between different samples, and therefore the distribution, every quantile, of intensity values should be *equal* for all samples [131,135]. One might argue that this could remove true differences between samples, especially in the tails of the distribution where a probe could get the same value on each array. This has been argued not to be a problem since multiple probes are used for each probe set, and individual probes would be scattered throughout the distribution [135]. Others have indicated that RMA reduced variance to inappropriate levels thus leading to high false positives [138].

When large differences are expected, as when comparing a stimulated cell line to an un-stimulated one, quantile normalization is improper since a multitude of genes are suspected to show differential expression. The effect of MAS 5, RMA and GC-RMA normalization can be viewed in Figure 4 showing these values for some of the arrays used in *Study III*, illustrating the large difference between them.

In *Study III* we chose to use the GC-RMA normalization, because it considers probe sequence and thus some on the bias due to probe effects can be eliminated.

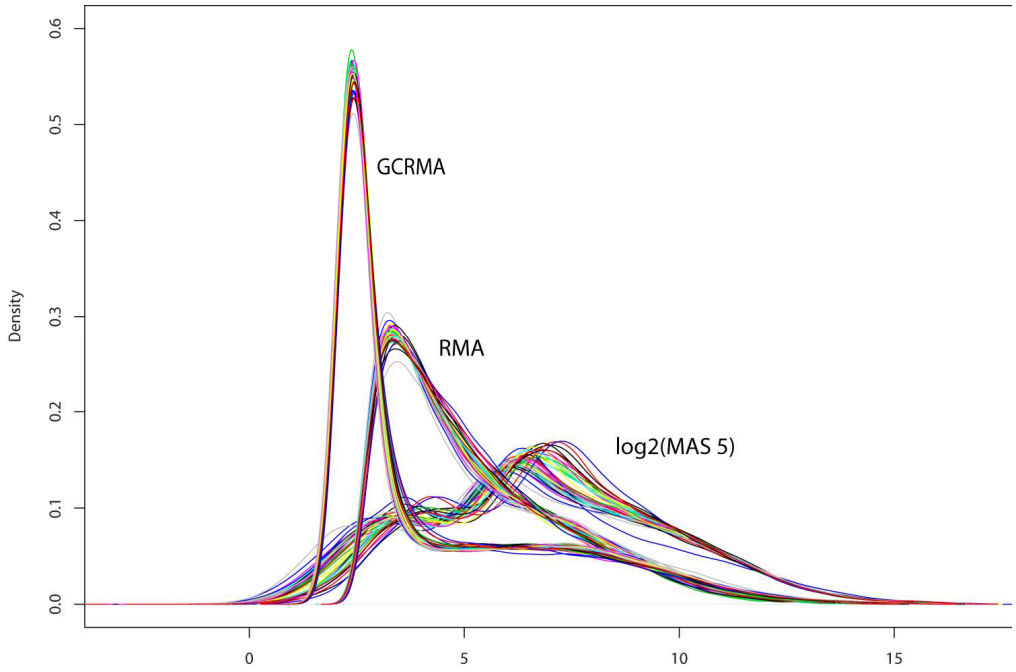


*Figure 4. Our data from Affymetrix arrays, measuring gene expression in CSF samples, shows intensity (x-axis) plotted against proportion of probe sets (y-axis). Data normalized by MAS 5 (plotted log2 transformed), RMA or GC-RMA.*

Affymetrix gene expression arrays has several basic quality measures: average background, scale factor, percent of genes called present, 3′ to 5′ ratio of GAPDH (glyceraldehyde-3-phosphate dehydrogenase) and β-actin and spike-in probes, these are intended to show how well the amplification, labeling and hybridization has proceeded. We investigated these measures using the Simpleaffy package [139] in R [70]. The 3′ to 5′ ratios are known to deteriorate when using a two cycle labeling procedure, which could be seen in our results as well. Three samples showed deviating values throughout Affymetrix quality control (QC): two PBL MS relapse and one control CSF, and were not included in the final analyses.

## 4.5 STATISTICAL ANALYSIS OF GENE EXPRESSION PROFILING DATA

Gene expression profiling produces a wealth of data, in *Study III* about 54,000 data points per sample and a maximum of 26 vs. 18 individuals in one comparison. Just by chance thousands of transcripts will be differentially expressed. It is logical to assume that in a given tissue at a given time (at sampling) not all genes are transcribed, and one might therefore want to remove data points produced by non-expressed genes prior to analysis in order to increase power (less probe sets to analyse, less correction for multiple testing). We removed probe sets with a low variance across samples in

*Study III*: the normalized (by GC-RMA) intensity values for a given included gene should have a reasonably large variance, as detected by an inter quantile range of at least 0.5. Undoubtedly truly expressed, and even truly differently expressed, probe sets might have been removed trough this filtering step. Given the decrease in numbers of analysed probe set from 54,000 to around 14,000, and the results from the proceeding analysis, this filtering step does not seem illogical and increased our power to detect differentially expressed probe set. Another route of filtering might be judged on the intensities of the probes: whether Affymetrix algorithm called the probe set present on a reasonable amount of arrays among others. I report all fold changes on the raw scale, even though values per probe set is on the log2 scale, where the mean for each group is calculated on the log2 scale. Thus, fold changes are reported as:

Fold change = $2^{(\text{mean (group 1 log2 values)} - \text{mean (group 2 log2 values)})}$

At the initiation of statistical analysis, data is retained in a gene expression data matrix with rows corresponding to probe sets and columns corresponding to samples.

### 4.5.1 Pattern discovery

Using unsupervised methods, one can try to detect structures in the data while not taking considerations to the sample or gene labels. Using clustering techniques objects are clustered based on measurements of similarity of expression, and one expect that functionally related genes and samples cluster together [140]. Several types of measurement of similarity exist, such as Euclidean: the geometric distance in the multidimensional space or Manhattan**:** average difference across dimensions. We used clustering of samples in *Study III* in order to support our notion of exclusion of low quality arrays; samples that showed deviant values in these initial quality assessments were also deviant in clustering and component analysis were therefore removed from further analysis (three samples mentioned in the earlier paragraph: two PBL MS relapse and one control CSF).

Several dimensional reduction techniques exist where similar entities are combined. Principal component analysis (PCA) is one of these methods, here linear combinations of the row or column vectors are computed, and the first principal component contains the largest variation within the samples, and the second one is orthogonal to

the first one and contains the second largest variance within the data, and so forth [141]. Thus, principal component one describes the largest variance, and hopefully the first two or three components will describe most of the variance in the data set since they are easily plotted against each other and visualised. In *Study III* PCA was conducted using the made4 package [142] and Figure 2 shows the resulting array projections.

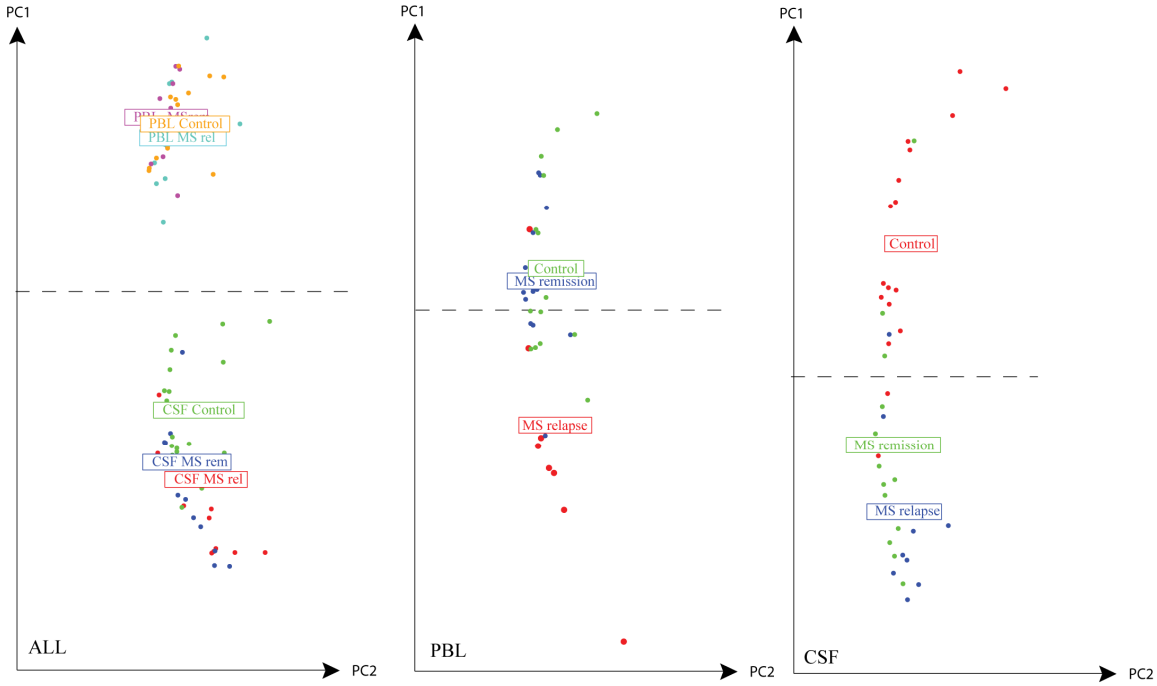


*Figure 5. Principal component representation of samples, plotting principal component one against principal component two.*

As expected probe sets that detect different expression levels depending on tissue examined created the first principal component. When the PCA is conducted separately in each tissue, relapse samples seem to separate from others among PBL samples, and control samples seem to separate from others in CSF samples.

### 4.5.2 Single transcripts

Usually one of the goals in a gene expression profiling study is to detect differently expressed genes between two or more conditions. Different test statistics do however rely on different assumptions regarding the distribution. In gene expression profiling numerous statistical methods are used to assess significant differential expression, and I will review some of the most commonly used.

A t-test is a simple statistical test designed to detect differences in means between two populations with values from a normal distribution [72]. The test statistic has the form:

$t = (mean_1 - mean_2) / (s * \sqrt{N})$

Where $s$ is the standard deviation and $N$ is the number of samples. Many of the most popular methods in gene expression profiling are t-tests or modified t-tests. A Welch t-test does not assume that the variance in the two groups is equal, and thus the denominator in the t-statistic is slightly changed ($\sqrt{(s_1^2/N_1 + s_2^2/N_2)}$). Dudoit et al. [141] proposed the use of permutations of sample labels to estimate the null distribution of the Welch t-statistic, thus avoiding any assumption on distribution. Genes or probe sets with low fold changes could also have lower variance, and can thus more easily be judged statistically significant differentially expressed. Therefore the significance analysis of microarrays (SAM) methodology adds a fudge factor to the denominator of the t-statistic where the fudge factor is calculated from the distribution of gene-specific standard errors [143]. Using the Limma package [144] the log-odds of differential gene expression is modelled by linear regression. A t-test with a Bayesian adjusted denominator is used, this has been described as equivalent to shrinkage of the estimated sample variances towards a pooled estimate, resulting in more stable results when the number of arrays is limited [145].

Different rank based methods exist as well, the rank product methods calculate the product of all ranks for every gene divided by the total number of examined genes, genes are thereafter ranked according to their rank product [146]. One can also assess significance by the area under a receiver operating characteristic (ROC) curve. Here true positives and false positives from a classification procedure are plotted against each other, and the area under the curve provides an estimate of the probability that the gene is regulated between the two groups [147].

When performing a single statistical test one might feel at ease with knowing that the probability of acquiring a significant p-value when the null hypothesis is in fact true is limited to 5 %. When considering multiple items at once, as during a microarray investigation, the equivalent action would produce 2,700 type I errors (for a 54,000 probe set array). Several means to adjust for this has evolved. One can change the level of the test in order to limit the chance of *one* false positive to a desired level, one such method is the Bonferroni correction and simply reduces the level of α by the number

of comparisons or independent comparisons. This is a conservative correction that will produce many type II errors. Another lines of action includes limiting the proportion of false positives among *all* test judged as significant. The false discovery rate (FDR) correction was first proposed by Benjamini and Hochberg [148]. Here the items are arranged in a list according to significance level, and the first item to survive the correction is the one with a p-value of: $p_i < (i/m) * q$

Where $i$ is the index in the ordered list, $m$ is the number of tests and $q$ is the desired FDR. Thus this methodology takes consideration to the number of tests as well as how many tests that initially were judged as significant. If 1000 tests are made and 50 tests are significant, no test will survive the correction (given that $\alpha = q = 5\%$). If however all 1000 test are judged significant, no correction will be performed since this indicates that the accepted number of false positives exist in the result already.

In *Study III* we chose to use a Welch t-test with a null distribution created by 1,000,000 permutations to detect differentially expressed probe sets. Thereafter we restrained the FDR to 5 % using the multtest package [149] in R [70].

Since we had an unbalanced study population regarding sex among our MS patients sampled in relapse or remission we performed a linear regression [144] on status as well as sex in order to elucidate regulation independent of sex in the PBL comparison.

### 4.5.3 Sets of transcripts

When performing a gene expression profiling project, one might become overwhelmed by a large number of differentially expressed genes from which hypothesis regarding disease processes should be formulated. Luckily for us, researchers have been studying genes and proteins and their functions and connections for a long time and have created tools for inferring functional properties of long lists of genes. Genes can be joined by their participation in certain pathways, available pathway databases include: Gene Ontology [150], Kyoto Encyclopedia of Genes and Genomes (KEGG) [151], Ingenuity Pathway Analysis (IPA, www.ingenuity.com, Ingenuity Systems) and several other commercial alternatives. The expression of genes can additionally be controlled by the same transcription factor (TF), information regarding that can be extracted from the TRANSFAC database [152,153].

After deciding which sets of genes to analyse, one needs to decide how to perform the analysis. One standard procedure involves determining enrichment of significantly differentially expressed genes within a pathway using a $\chi^2$-test or Fisher's exact test, and this is the procedure used in IPA. In *Study III* the differently expressed genes from the CSF comparison of MS and control samples as well as those from the PBL comparison of relapse vs. remission were analysed using IPA.

Mootha et al. [154] suggested using a method not drawing the arbitrarily line between significantly and not significantly differentially expressed genes. Instead subtle but coordinated changes in expression of genes belonging to the same gene set can be assessed. Here the distribution of genes belonging to one pathway in the full list of sorted genes is assessed, and clustering of genes on this list indicates coordinated differential expression. We have used a similar methodology in *Study III*, where the t-statistics for each gene belonging to a KEGG pathway are joined and normalized and compared to a null distribution created by permutation, as implemented in the Category package [155] in R [70]. Thereby we assessed the regulation within predefined sets of genes using two different databases (KEGG and IPA) and two different statistical approaches.

### 4.5.4 Networks

Building networks is a way of characterizing your data, and is meant to illustrate contexts within your data that would otherwise go unnoticed. A network can either be built bases solely on the actual data or with information from other sources as well. In *Study III* we built networks in IPA using their database on published connections between molecules, e.g. protein-protein binding or influence of expression. Thus, possible functional networks can be built that are not part of settled pathways, and could thus shed light on potential functional groups in the investigated traits. IPA only displays small networks with a maximum size of about 35 molecules, and the highest rated network is the one with most connections. The reason for this restriction is that large complex networks are hard to grasp and draw conclusions from.

### 4.5.5 Annotation

As mentioned earlier the probes on Affymetrix arrays were designed to hybridize to RNA sequences found at the time in the UniGene, bdEST, GeneBank and RefSeq databases. The content of such databases are constantly developed, refined and rearranged. Therefore, some of the probe sets found on the arrays does not align to any known sequence today, and those that do can be given different annotations due to different strategies of annotation. Affymetrix builds their annotation on UniGene clustering, which is a fully automated process where sequences that align around the same gene or each other are joined into clusters that automatically get an annotation. In *Study III* we used Affymetrix annotation through the Bioconductor package hgu133plus2.db [156] and realized that this route of annotation creates some discrepancies, especially since a number of probe sets aligning to immunoglobulin transcripts were annotated as *HLA*-C when using the Affymetrix annotations. Thus we sought another route of annotation that demanded that the probe sets aligned to well-characterized mRNA sequences. Ensembl's Biomart is an attempt to correlate different biological databases to each other. Using the biomaRt package [157] we could therefore in *Study III* collect those Ensembl transcripts that aligned to our probe sets, and from these we also collected RefSeq ID's and HGNC (HUGO gene nomenclature committee) gene symbols as well. This more conservative method does not annotate a rather large proportion of the probe sets, and we therefore decided to use both methods. There was however a small number of analyzed probe sets (6-7%) that did not acquire annotation using any of these methods even though they seemingly are detecting expressed genes.

The knowledge on sequences grow for each day, and probe sets designed a couple of years ago will most certainly not be the ultimate answer. Approaches have been proposed where alternative mapping of probes of Affymetrix arrays are based on up-to-date knowledge [158,159].

## 4.6 CONFIRMATION OF DIFFERENTIAL EXPRESSION

There is a multidimensional problem with most gene expression profiling studies; since there are many times more probe sets than samples to be compared. Due to this and occasionally additional problems, validation of the observed differential expression may be warranted. One could either perform a technical confirmation;

detect the same differential expression using the same samples again, or a biological confirmation using new independent samples. Confirmational studies are usually performed using quantitative real time PCR (qRT-PCR).

### 4.6.1 Quantitative real time PCR

qRT-PCR is a technique whereby the amount of transcript (absolute or relative to input of mRNA or another gene) is detected in real time during a PCR reaction. The logarithmic increase in target is detected using different probes or dyes.

Real time PCR has three phases: exponential phase, linear phase and plateau phase. In the exponential phase the product increase exponentially, doubling in each cycle if the efficacy is at 100 %, since there is no limitation in reagents. As reagents become limited a more linear increase in product is seen, which eventually, as some reagent become depleted, reaches a plateau. Plotting logarithm 2 transformed product (as measured by fluorescence of probes) versus cycle number yields a linear range at where signal correlates to with the original template [160,161]. A horizontal threshold is set above background disturbances during the exponential growth, and the number of cycle it takes for a product to reach this threshold is denoted the Ct value and functions as the primary statistic in RT-PCR. Quantity of target can be assessed either absolutely or relatively where the absolute quantification employs a calibration curve in order to derive the input of target. Relative quantification estimates quantity of target in relation to an internal reference gene, which in turn is compared to target vs. reference in other samples. Usually the exact copy number of target is not of primary importance, thus relative quantification is employed. Several data analysis method exist in relative quantification, two of the most common are the efficiency calibrated method [162,163] and the ΔΔCt method [164], and both assume that the efficiency of amplifications is the same across all samples. The experiment will involve serial dilutions of samples for PCR detection of both target and reference gene, and the Ct number is plotted against cDNA input and the slope is used to calculate the amplification efficiency. ΔCt is calculated by subtracting the Ct value of target from the reference, and ΔΔCt is then calculated by subtracting a control ΔCt value, i.e. the mean ΔCt value among controls, from each ΔCt value. If the amplification efficiency is good, around 100 %, the expression ratio equals $2^{-\Delta\Delta Ct}$.

One can also calculate the expression ratio while taking consideration to difference in efficiency between reference and target:

$$(efficiency\ target)^{\Delta Ct\ (target)} / (efficiency\ reference)^{\Delta Ct\ (reference)}$$

Another method compares the individual Ct values against respective standard curve, and expression ratio is calculated as the ratio between these two measures [165]. This methodology can also be used to assess absolute quantification if the exact quantities in the standard curves are known.

As discussed by others [166] the correlation between microarray data and RT-PCR data is quite poor, especially for low fold changes. Variability in technical procedures affects both methods and in particular the route of normalization differs greatly. The choice of reference gene in qRT-PCR is critical, and should have a stable non variable expression in all samples investigated [167], moreover the utilization of several reference genes simultaneously is more stable [168]. When analysing replicated samples, separated just prior to PCR, the difference in Ct values becomes approximately 0.5 Ct. Including all further handling of samples the variation will be greater, and differences of one Ct (which equals a fold change of 2) will usually not be detected (personal communication, Per Larshammar, Applied Biosystems).

In *Study III* we used a SYBR green detection system (Bio-Rad) and the ΔΔCt method together with a nonparametric Wilcoxon signed-rank test to validate differential expression. We attempted at confirming the regulation using the same samples as those used in the original gene expression profiling, and additionally added an independent population of MS patients and controls. The groups remained quite small, and thus they were joined in the final analysis. The primers for the quantitative PCR were designed to bind to the same sequences as the probes on the Affymetrix array. Five of the most significantly differently expressed transcripts were selected for validation in the CSF of MS patients and controls, using *GAPDH* as a reference gene. All these genes displayed large fold changes, and all were confirmed as differentially expressed. We also tried to validate ten transcripts in the PBL comparison of MS patients in relapse to those in remission, these transcripts showed fold changes ranging from 0.58 to 2.29. Initially *GAPDH* was used as a reference gene, and one of the investigated transcripts was confirmed (one additional transcripts displayed a p-value of 0.06). When investigating the variance of intensity for probe sets detecting

*GAPDH* in our Affymetrix data, we realised that *GAPDH* seemed to vary more than the investigated transcripts. We therefore selected a new reference genes based on our microarray data: *TGS1* (trimethylguanosine synthase homolog). The correlation between *GAPDH* and *TGS1* was relatively high: 0.909, but *TGS1* had a much lower expression. We were however not able to confirm differential expression of any of the investigated transcripts. It is well known that confirmation of differential expression of fold changes below two is uncommon [166] and depends on concentration of target and qRT-PCR method used [169]. The use of several stable reference genes averaged together with a geometric mean would give a more stable normalization [168 ], and thus the ability to detect smaller fold changes. But the use of the geometric mean of *GAPDH* and *TGS1* as reference did not validate any transcripts in *Study III*. The nature of differently expressed probe sets in the two comparisons did however differ more than in the magnitude of fold change; differently expressed probe sets during disease bouts in PBL did not align well with validated ensembl transcripts. Out of the 1031 significant probe sets in PBL 61 % did not map to an Ensembl transcript whilst only 28 % of the 4176 significant probe sets in the CSF comparison of MS patients to controls lacked this conservative annotation. This indicates that transcripts other than the most validated and characterized are differently expressed during an MS bout. The qRT-PCR data did however showed some tendencies for association, and these trends followed the initial findings, thus a larger study population might have validated additional transcripts.

Could the lack of confirmation in our PBL comparison indicate that no true differential expression is present there? We have performed a rather assumption free significance testing based on permutation and a conservative correction for multiple testing, and still more than a thousand probe sets were judged as differentially expressed. It could be that we were unlucky in our choice of transcripts to validate, and that those ten transcripts are among the 50 (on average) false positives. I think that a more likely explanation is that qRT-PCR is not ideal for detecting these small differences and that using a larger population and optimising the procedure might have validated more transcripts. Another explanation is that the differential expression seen is merely a consequence of a systematic bias, but the nature of such a bias is unknown to us.

## 4.7 SUMMARY OF RESULTS OF STUDY III

### 4.7.1 Comparing MS patients and controls

We detected over 4,000 differently expressed transcripts in CSF cells comparing MS patients (sampled both in relapse and remission) to controls, whereas the same comparison in PBL did not show any differentially expressed transcript. This was also illustrated using PCA where CSF samples from MS patients clustered distant from a cluster of controls (Figure 5). This implies that CSF cells demonstrate some aspects of the disease processes in MS patients, in contrast to PBL. The most differentially expressed probe sets were immunoglobulins that showed fold changes of up to 500, which act as a positive control since we demanded oligoclonal IgG bands in all assessed MS patients. Five of the most differently expressed probe sets were also validated using quantitative RT-PCR: *AIF1* (allograft inflammatory factor 1), *TNFRSF17* (tumour necrosis factor receptor superfamily, member 17), *MGC29506*, *POU2AF1* (POU class 2 associating factor 1) and *PLAUR* (plasminogen activator, urokinase receptor).

IPA analysis in CSF revealed 45 significantly enriched pathways, where "TREM1 Signalling" and "Role of Pattern Recognition Receptors in Recognition of Bacteria and Viruses" were the two most significant. Overall immune-related pathways dominated the list of enriched pathways. The analysis of coordinated regulation within KEGG pathways in CSF illuminated the roles of "Complement and coagulation cascades" as most significantly downregulated in MS, and "Cell cycle" as most significantly upregulated in MS. Overall, this analysis illuminated the roles of protein export and degradation, basal transcription factors and other fundamental processes.

Networks were built based on our data as well as known connections between molecules contained in the knowledge database of IPA, and these were scored based on their connectivity. The highest scoring network represented biological functions related to protein degradation, protein synthesis, and cellular assembly and organization. One of the hubs in this network is LMNA (lamin A/C), a protein in the nuclear lamina protein network underlying the inner nuclear membrane that determines nuclear shape and size, and which is involved in cellular integrity and gene expression [170]. CTNNB1 (catenin (cadherin-associated protein), beta 1), an adherens junction protein, makes up another subnode. The second highest scoring cluster is centred around CDKN1A (cyclin-dependent kinase inhibitor 1A), a regulator of cell

cycle progress at G1, and the functional analysis showed enrichment of molecules involved in "cell cycle", "cancer" and "reproductive system disease". The third most highly ranked network contains molecules involved in "immune and lymphatic system development and function", "tissue morphology", and "haematological system development and function".

Transcription factor binding site analysis showed enrichment of 84 TF binding sites, most significant were the E2F family and ZFP161 (homologue of zink finger protein 161, mouse).

Although no single transcript was deemed differentially expressed in PBL, the analysis of coordinated regulation of KEGG pathways suggest the importance of 41 pathways, most significant being "ECM-receptor interaction" detected as upregulated in MS, and "Citrate cycle (TCA cycle)", downregulated in MS.

### 4.7.2 Comparing MS patients in relapse and remission

The most striking among our findings was the suggestion that a disease bout is companioned by a substantial difference of transcription in PBL, albeit no simultaneous differential expression can bee seen in the CSF. The 1031 differentially expressed probe sets contain a surprisingly large proportion of probe sets lacking alignment with Ensembl transcripts, over 60 %, whilst the entire list of analysed probe sets contain 37.7 % such probe sets. This implies that quite uncharacterized transcripts are differentially expressed in the PBL of patients with relapse.

The IPA analysis suggested the importance of 43 pathways, the two most significant being "SAPK/JNK Signalling" and "Hypoxia Signalling in the Cardiovascular System". The analysis of coordinated regulation within KEGG pathways showed 60 pathways upregulated during a relapse, and 14 downregulated. Noticeable is that upregulated pathways consisted of pathways related to development, metabolism and basic cell signalling, such as "Dorso-ventral axis formation", "GnRH signalling pathway", "Fatty acid metabolism" and "Notch signalling pathway", whereas a majority of downregulated pathways were connected to the metabolism of several amino acids.

The highest scoring network contained molecules involved in "gene expression", "protein degradation" and "protein synthesis". The mediator complex (also known as TRAP, SMCC, DRIP, or ARC) and other molecules involved during transcription formed one small highly interconnected subcluster, whereas two transcription factors (SP1 and VHL (von Hippel-Lindau tumour suppressor)) formed two additional nodes. The second highest scoring network was related to "cancer", "reproductive system disease" and "neurological disease", and one major node was the signal transduction protein YWHAZ (tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, Z isoform). The third most highly ranked networks major node was the tumour suppressor PTEN (phosphatase and tensin homolog) and related functions were "gene expression", "DNA replication, recombination and repair", and "cell cycle".

The TF binding site analysis implied 29 enriched binding sites, most significant were binding sites for Hb (Hunchback) and PLZF (Promyelocytic leukemia zinc finger, also called zinc finger and BTB domain containing 16 (ZBTB16)).

### 4.8 DISCUSSION ON MS TRANSCRIPTOMICS TODAY

To our knowledge, we have performed the first gene expression profiling experiment investigating MS processes using samples from CSF, and additionally we also included paired samples from PBL. Thus we could detect large differences in the regulation of transcription within these two tissues, which suggested that even though CSF of MS patients differ greatly in expression of genes as compared to controls, a disease bout could only be detected when assessing the gene expression in PBL.

Large scale gene expression in CSF has not been performed in MS previously, supposedly due to the demands on large amount of RNA, thus these results can not be discussed while considering other gene expression profiling studies. We measured gene expression in the total cell population of CSF from MS patients and controls with other neurological diseases. Thus, we expected samples consisting of about 80-90 % T-cells, and a few percent of NK-cells, NK-like T-cells, B-cells, plasma cells and monocytes, where MS patients would have a smaller proportion of monocytes and NK-like T-cells, and a larger proportion of B-cells and plasma cells [108]. We did see many upregulated transcripts connected to B-cells, e.g. *TNFRSF17* (tumour necrosis factor receptor superfamily, member 17, also called B-cell maturation factor, BCMA), *POU2AF1*, *MS4A1*

(membrane-spanning 4-domains, subfamily A, member 1) and *CD24* which were upregulated by fold changes of 20, 15, 4 and 5 respectively in MS CSF as compared to controls. The most differently expressed transcripts were connected to immunoglobulin with fold changes of up to 500. These differences might in part reflect the expected increase of B cells (7 fold) and plasma cells (13 fold) but not entirely. In the pathway analysis we saw marked decrease of pathways involved in monocytes and macrophages such as "TREM1 signalling" and "Complement and Coagulation cascades". A very limited number of studies have investigated gene expression in the CSF of MS patients, but we do confirm lower expression of *VEGF* [171] in MS patients compared with controls.

A number of transcripts earlier found to be upregulated within MS lesions, were suggested to be downregulated in the CSF of MS patients in *Study III*: *TREM2*, *C3* and *C1qB* with fold changes of 0.17-0.37 [172 ], *ALOX5* with fold change 0.43 [173], *IL18*, *IL1β* and *CCR1* with fold changes of 0.23-0.53 [174]. *PLAUR* was also expressed proportionally less in MS CSF as compared to controls in our study, but have been shown to have a higher protein expression in lesions and normal appearing white matter of MS patients [175]. Moreover the PLAU-complex has unique abilities to facilitate transmigration trough the BBB [176]. These changes might reflect the selective transmigration of cells from CSF further into CNS.

Additionally, various results indicate more proliferating and active cells in the CSF of MS patients: ribosomal genes, cell cycle genes, "cell cycle" pathway, "protein export" pathway and "basal transcription factors" pathway were all upregulated in the CSF of MS patients.

When comparing the expression in PBL from MS patients and controls, two KEGG pathways involving integrins were upregulated: "ECM-receptor interaction" and "Focal adhesion". This might implicate that migration and factors enabling migration to the CNS is increased in peripheral blood of MS patients.

The large amount of genes differently expressed between samples collected during a relapse or remission in PBL were enriched for inflammatory and stress related pathways, and gene expression, protein degradation and protein synthesis were also

distinguished. One must bear in mind, however, that all pathway analyses are built on the annotation, which was troublesome for the differentially expressed probe sets in this comparison (60 % lacked the more conservative annotation). It's hard to draw conclusions regarding the ongoing processes when the involved transcripts are unknown, and further attempt to characterize these transcripts would be useful. However, the UniGene based annotation does imply what gene or at least which genetic region (close to the gene) that is involved, and therefore the following analyses do reveal important knowledge. We are eager to see coming studies investigating the gene expression in relapse/remission and whether our findings will be validated.

Some other studies have investigated gene expression in cells from blood and compared patients in relapse to remission or controls. Arthur et al. [117] investigated patients with MS sampled during a disease bout (n = 10), or during remission (n = 10), and blood donor controls (n = 6, 20 RNA samples from women were pooled prior to microarray procedure, 5 samples from males were used individually). Gene expression was examined in whole blood, not lymphocytes, and this is the most likely explanation to the disagreement in results. No correction for multiple testing was conducted in the Arthur et al. study, but the most noticeable is that the regulation seemed to be as big in relapse as in remission, in contrary to our results. Arthur et al. focused especially on *ALOX5* which they found upregulated in both relapse and remission as compared to controls [117]. We confirm the upregulation of *ALOX5* in relapse samples using three probe sets, with fold changes of 1.5 to 2.4. In remission however, the *ALOX5* probe sets were rather downregulated (fold change of 0.7) in our material but not significant even when judging from the uncorrected p-value. Others have examined gene expression in peripheral blood mononuclear cells with special emphasis on apoptosis related gene expression and relapse [119]. Here many (1,578) differentially expressed genes were found comparing patients in relapse to those in remission, and genes relevant for regulation of apoptosis, caspase activity, and caspase regulation activity were significantly downregulated in acute relapse. We confirm the regulation of a number of genes: *BCLAF1*, *TAX1BP1*, *RTN4* and *SMNDC1* among others, and moreover we also see enrichment and regulation of apoptosis related pathways such as "Apoptosis signalling" and "Death Receptor Signalling".

Simultaneously with the regulation seen in PBL, no differential expression is seen in CSF between relapse and remission samples. The KEGG analysis reveals only upregulated pathways during a disease bout which were related to basic metabolism, maybe indicating somewhat more active cells.

Gene expression profiling in MS is still in its infancy, although all conducted studies do reveal part of ongoing processes. Many more diverse investigations are needed to firmly establish the ongoing pathological processes.

# 5 CONCLUDING REMARKS

This thesis is focused on genetic and transcriptomic studies of MS, both methodologies promising to reveal parts of the MS aetiology. It is undoubtedly so that other areas of research also are needed in order to discover all risk factors for developing MS, such as epigenetics and epidemiological studies of environmental risk factors. We have however been able to investigate the relationship between twelve risk factors in this thesis: female sex, *HLA-DRB1, HLA-A, IL7R, IL2Rα, CLEC16A, CD58, RPL5, FAM69A, CD226*, *SH2B3* and *KIF1B*. This alone demonstrates the huge progress made within the MS research field in the last couple of years, where the studies in this thesis played a minor role. The risks these factors confer are not caused by confounding of each other, but several epistatic effects between them were suggested and deserve further investigation.

*CD58* and *RPL5* were investigated genetically partly because of their differential expression in the CSF of MS patients as compared to controls. Additionally *RPL5* was shown to be expressed at a lower level during a disease bout as compared to remission in PBL. Intriguingly both *RPL5* and *CD58* have been suggested to have connections to the differential expression of p53 family members [177,178], and the "p53 signalling pathway" was detected as enriched of differentially expressed probe sets during disease bouts. Other MS susceptibility genes were also suggested to be differently expressed in *Study III*; *KIF1B* was upregulated in PBL during a disease bout, and downregulated in CSF of MS patients and *SH2B3* was downregulated in CSF of MS patients. A number of HLA transcripts were differently expressed in the CSF of MS patients, mostly class II molecules showing a lesser expression. A limited number of the individuals included in the gene expression profiling were investigated genetically, thus no genotype phenotype correlation analysis has been performed, but this is a rational next step.

If I were given funds and additional time, I believe it would be worthwhile to investigate the implicated transcription factors in *Study III* genetically. Moreover, those genes with differential expression as well as genetic association could also be assessed more thoroughly. I expect that additional studies on tissues from MS patients will be performed, and hopefully able to validate the results presented here. It is my hope that

our gene expression profiling data, now retained in a public repository, can aid the researchers in our group as well as others.

As mentioned above, interdisciplinary research across many fields will be necessary to understand MS aetiology. But there is also great deal to be accomplished in each individual field. Through the rapid development of whole genome resequencing the utilization of such data is not far off, so we should prepare ourselves in order to be able to take advantage of that technical breakthrough. When whole genome resequencing has become the standard procedure the existence of each independent genetic risk factor might eventually be known, and gene-gene interactions investigated. However, at that point functional understanding needs to be acquired as well, an equally tedious task. One utopian study would involve whole genome resequencing data as well as transcriptional and proteomic profiling from many tissues at many time points from thousands of patients and controls. It might take some time, but it is likely that we will eventually get there. Meanwhile, progress is rapid and I am sure that our knowledge on MS aetiology will continue to grow, and that all pieces of the puzzle can lead to a better understanding of the disease and most importantly, ultimately help individuals suffering from MS.

# 6 ACKNOWLEDGEMENTS

First of all I would like to thank my supervisor Professor Jan Hillert, for answering that first letter and inviting me. You gave me the means and let me roam free, chasing my dreams and fighting to reach my goals. Your door was always opened, and you encouraged and supported me. Therefore these years have been challenging, educating, creative and filled with lots of fun.

I have a lot to thank my co-supervisor Dr. Kristina Duvefelt for. You have been there since day one, and it felt natural and easy to work with you. During the last couple of years your field of responsibility has grown tremendously but still you always have time to answer my questions and giving me advice. Thank you for believing in me, all the "harsh" discussions, the chocolate, the laughs, and the collaboration. A special big thank you for reading my thesis during your vacation…

Many thanks to the best colleagues in the world! We have had the most intense discussions about interaction, confounding and life after death. I have met new friends and had a blast. Thanks to all members at the Division of Neurology, especially:
☻ Iza Lima: Thank you for the constant chatting in our room about life and statistics, having a different point of view, our fruitful collaborations and reading my thesis!
☻ Kerstin Imrell: For always finding the weak spots in my argumentations, questioning everything and thereby making our group what we are (the smartest bunch ever!).
☻ Jenny Link: For great skiing companion, crayfish parties, HLA discussions, reading my thesis and collaboration that will continue!
☻ Jenny Ahlqvist: Thank you for lab-parties (very few since you left..!), a great holiday in Atlanta and help with planning my future life. Lets become neighbours soon!
☻ Frida Lundmark: For all those slopes "we" conquered in the Rockies, and always having the time to chat about life.
☻ Wangko Lundström: For being able to melt in to our, at that time, quite homogeneous group, and adding your wits and cocking skills!
Thomas Masterman, Eva Greiner, Ingrid Kockum, Eva Lindström, Viriginja Karrenbauer, Rasmus Gustafsson, Malin Lundkvist, Anna Fogdell-Hahn, Ajith Sominanda, Mathula Thangarajh, Leszek Stawiarz, Sebastian Yakisich, Kosta Kostulas, Christina Sjöstrand, Kristina Gottberg, Sverker Johansson, Lena von Kock, Lotta Widén Holmqvist, Andreia Gomes, Marina Vita, Urus Rot, Yassir Hussein, Susanna Mjörnheim and all others that have slipped my mind at this point. Thanks also to all summer students, project workers and master thesis students.

For all help with practical matters: Gunnel Larsson, Yvonne Sjölind, Cecilia Svarén Quiding, Merja Kanvera, Inge Gerd Löfving, Faezeh Vejdani and Anna Mattsson. Many thanks also to all nurses at the outpatient-clinic for help with sampling from the patients.

Thanks for all the fika and lunch company Eva, Annelie and Marjan. An additional thank you to Marjan for help with arranging food for the party!

My co authors at CMM: Tomas Olsson, Mohsen Khademi and Erik Wallström.
The gang at CMM plan 4, for lab-parties and other social activities, especially Ame, Allan, Melanie, Maja and Pernilla.

My co authors in Oslo: Cathrin Smestad, Åslaug Lorentzen, Anne Spurkland, Benedicte Lie, Elisabeth Celius and Hanne Harbo, it was a pleasure working with you!

My co authors at MEB: Juni Palmgren and Gudrun Jonasdottir Bergman. A special thank you to Gudrun for reading my thesis and all those long discussions about logistic regression and life.

Rebecca Ceder: for all the discussions about microarray analysis, for becoming a friend and on top of that reading Paper III!

The organizers and attendees at all those great courses at CSHL, KTH, Duke and KI.

All members of the Nordic MS genetics group for collaborations and interesting meetings.

Personnel at the Mutation Analysis Facility for genotypes.

A warm thank you to my friends, past and present, who might not have seen much of me lately. You mean so much to me. A special thank you to Linda & Johan, Malin & Per, Helena, Ann, Karin, Frida & Anders and Hansson. A special thank you to Linda for helping out with the party!

I would like to thank the new members of my family: Elisabeth, Jörgen, Birgitta, Erik & Cecilia, Valle & Anders.

To my family: Ragnar, Karin, Lena, Sara and Alexandra. To my mother and father because they never quite understood what I was doing, but were proud of me even so. Many thanks to my father also for all hours spent in our apartment, with a really great result! My sisters Lena and Sara, you made Stockholm my new home by opening your homes, for all the fun we have had, for being so close. My brothers in law Marcus and Fredde. Many thanks also to my nephews Algot and Hampus for showing me what life is really about.

Johan, life with you is the life I want to live.

# 7 REFERENCES


1. Compston A, Werkerle H (2005) The genetics of multiple sclerosis. In: Compston A, Confavreux C, Lassman H, McDonald WI, Miller D et al., editors. McAlpine's Multiple Sclerosis. 4 ed. London: Elsevier Inc.
2. Jersild C (1978) Studies of HLA antigens in multiple sclerosis. Boll Ist Sieroter Milan 56: 516-530.
3. McDonald WI, Compston A, Edan G, Goodkin D, Hartung HP, et al. (2001) Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. Ann Neurol 50: 121-127.
4. Miller DH, Albert PS, Barkhof F, Francis G, Frank JA, et al. (1996) Guidelines for the use of magnetic resonance techniques in monitoring the treatment of multiple sclerosis. US National MS Society Task Force. Ann Neurol 39: 6-16.
5. Bermel RA, Bakshi R (2006) The measurement and clinical relevance of brain atrophy in multiple sclerosis. Lancet Neurol 5: 158-170.
6. Kurtzke JF (1983) Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology 33: 1444-1452.
7. McDonald WI, Compston A (2003) The symptoms and signs of multiple sclerosis. In: Compston A, Confavreux C, Lassman H, McDonald WI, Miller D et al., editors. McAlpine's Multiple Sclerosis. London: Elsevier Inc. pp. 287-346.
8. Roxburgh RH, Seaman SR, Masterman T, Hensiek AE, Sawcer SJ, et al. (2005) Multiple Sclerosis Severity Score: using disability and disease duration to rate disease severity. Neurology 64: 1144-1151.
9. Confavreux C, Compston A (2003) The natural history of multiple sclerosis. In: Compston A, Confavreux C, Lassman H, McDonald WI, Miller D et al., editors. McAlpine's Multiple Sclerosis. 4 ed. London: Elsevier Inc. pp. 183-269.
10. Greenberg BM, Calabresi PA (2008) Future research directions in multiple sclerosis therapies. Semin Neurol 28: 121-127.
11. Lassman H, Wekerle H (2003) The pathology of multiple sclerosis. In: Compston A, Confavreux C, Lassman H, McDonald WI, Miller D et al., editors. McAlpine's Multiple Sclerosis. 4 ed. London: Elsevier Inc. pp. 557-599.
12. Chitnis T (2007) The role of CD4 T cells in the pathogenesis of multiple sclerosis. Int Rev Neurobiol 79: 43-72.
13. Delgado S, Sheremata WA (2006) The role of CD4+ T-cells in the development of MS. Neurol Res 28: 245-249.
14. Martola J, Stawiarz L, Fredrikson S, Hillert J, Bergstrom J, et al. (2007) Progression of non-age-related callosal brain atrophy in multiple sclerosis: a 9-year longitudinal MRI study representing four decades of disease development. J Neurol Neurosurg Psychiatry 78: 375-380.
15. Trapp BD, Nave KA (2008) Multiple sclerosis: an immune or neurodegenerative disorder? Annu Rev Neurosci 31: 247-269.
16. Lassmann H (2007) Multiple sclerosis: is there neurodegeneration independent from inflammation? J Neurol Sci 259: 3-6.
17. Libbey JE, McCoy LL, Fujinami RS (2007) Molecular mimicry in multiple sclerosis. Int Rev Neurobiol 79: 127-147.
18. Compston A, Confavreux C (2003) The distribution of multipls sclerosis. In: Compston A, Confavreux C, Lassman H, McDonald WI, Miller D et al., editors. McAlpine's Multiple Sclerosis. London: Elsevier Inc.
19. Ascherio A, Munger KL (2007) Environmental risk factors for multiple sclerosis. Part II: Noninfectious factors. Ann Neurol 61: 504-513.
20. Ascherio A, Munger KL (2007) Environmental risk factors for multiple sclerosis. Part I: the role of infection. Ann Neurol 61: 288-299.
21. Hernan MA, Olek MJ, Ascherio A (1999) Geographic variation of MS incidence in two prospective studies of US women. Neurology 53: 1711-1718.
22. Wallin MT, Page WF, Kurtzke JF (2004) Multiple sclerosis in US veterans of the Vietnam era and later military service: race, sex, and geography. Ann Neurol 55: 65-71.
23. Consortium TIH (2003) The International HapMap Project. Nature 426: 789-796.

24. Rothman K (2002) Epidemiology : an introduction. New York: Oxford University Press.
25. Kissmeyer-Nielsen F, Svejgaard A, Hauge M (1968) Genetics of the human HL-A transplantation system. Nature 219: 1116-1119.
26. Jersild C, Fog T (1972) Histocompatibility (HL-A) antigens associated with multiple sclerosis. Acta Neurol Scand Suppl 51: 377.
27. Jersild C, Fog T, Hansen GS, Thomsen M, Svejgaard A, et al. (1973) Histocompatibility determinants in multiple sclerosis, with special reference to clinical course. Lancet 2: 1221-1225.
28. Hillert J, Olerup O (1993) Multiple sclerosis is associated with genes within or close to the HLA-DR-DQ subregion on a normal DR15,DQ6,Dw2 haplotype. Neurology 43: 163-168.
29. Fogdell-Hahn A, Ligers A, Gronning M, Hillert J, Olerup O (2000) Multiple sclerosis: a modifying influence of HLA class I genes in an HLA class II associated autoimmune disease. Tissue Antigens 55: 140-148.
30. Harbo HF, Lie BA, Sawcer S, Celius EG, Dai KZ, et al. (2004) Genes in the HLA class I region may contribute to the HLA class II-associated genetic susceptibility to multiple sclerosis. Tissue Antigens 63: 237-247.
31. The Games Collaborative G, Ban M, Booth D, Heard R, Stewart G, et al. (2006) Linkage disequilibrium screening for multiple sclerosis implicates JAG1 and POU2AF1 as susceptibility genes in Europeans. J Neuroimmunol 179: 108-116.
32. WTCCC (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661-678.
33. IMSGC, Hafler DA, Compston A, Sawcer S, Lander ES, et al. (2007) Risk alleles for multiple sclerosis identified by a genomewide study. N Engl J Med 357: 851-862.
34. Lundmark F, Duvefelt K, Iacobaeus E, Kockum I, Wallstrom E, et al. (2007) Variation in interleukin 7 receptor alpha chain (IL7R) influences risk of multiple sclerosis. Nat Genet 39: 1108-1113.
35. Gregory SG, Schmidt S, Seth P, Oksenberg JR, Hart J, et al. (2007) Interleukin 7 receptor alpha chain (IL7R) shows allelic and functional association with multiple sclerosis. Nat Genet 39: 1083-1091.
36. Zhang Z, Duvefelt K, Svensson F, Masterman T, Jonasdottir G, et al. (2005) Two genes encoding immune-regulatory molecules (LAG3 and IL7R) confer susceptibility to multiple sclerosis. Genes Immun 6: 145-152.
37. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, et al. (2007) Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. Nat Genet 39: 1329-1337.
38. Rubio JP, Stankovich J, Field J, Tubridy N, Marriott M, et al. (2008) Replication of KIAA0350, IL2RA, RPL5 and CD58 as multiple sclerosis susceptibility genes in Australians. Genes Immun.
39. Hoppenbrouwers IA, Aulchenko YS, Ebers GC, Ramagopalan SV, Oostra BA, et al. (2008) EVI5 is a risk gene for multiple sclerosis. Genes Immun 9: 334-337.
40. Weber F, Fontaine B, Cournu-Rebeix I, Kroner A, Knop M, et al. (2008) IL2RA and IL7RA genes confer susceptibility for multiple sclerosis in two independent European populations. Genes Immun 9: 259-263.
41. Aulchenko YS, Hoppenbrouwers IA, Ramagopalan SV, Broer L, Jafari N, et al. (2008) Genetic variation in the KIF1B locus influences susceptibility to multiple sclerosis. Nat Genet 40: 1402-1403.
42. Baranzini SE, Wang J, Gibson RA, Galwey N, Naegelin Y, et al. (2008) Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. Hum Mol Genet.
43. IMSGC (2008) The expanding genetic overlap between multiple sclerosis and type I diabetes. Genes Immun.
44. Hafler JP, Maier LM, Cooper JD, Plagnol V, Hinks A, et al. (2008) CD226 Gly307Ser association with multiple autoimmune diseases. Genes Immun.
45. Poser CM, Paty DW, Scheinberg L, McDonald WI, Davis FA, et al. (1983) New diagnostic criteria for multiple sclerosis: guidelines for research protocols. Ann Neurol 13: 227-231.
46. Kwok PY (2001) Methods for genotyping single nucleotide polymorphisms. Annu Rev Genomics Hum Genet 2: 235-258.

47. Kim S, Misra A (2007) SNP genotyping: technologies and biomedical applications. Annu Rev Biomed Eng 9: 289-320.
48. Olerup O, Zetterquist H (1992) HLA-DR typing by PCR amplification with sequence-specific primers (PCR-SSP) in 2 hours: an alternative to serological DR typing in clinical practice including donor-recipient matching in cadaveric transplantation. Tissue Antigens 39: 225-235.
49. Excoffier L (2003) Analysis of population subdivision. In: Balding D, Bishop M, Cannings C, editors. Handbook of Statistical Genetics. West Sussex: John Wiley & Sons, Ltd. pp. 713-750.
50. Hudson R (2003) Linkage disequilibrium and recombination. In: Balding D, Bishop M, Cannings C, editors. Handbook of Statistical Genetics. West Sussex: John Wiley & Sons, Ltd. pp. 662-680.
51. Xiong M, Guo SW (1997) Fine-scale genetic mapping based on linkage disequilibrium: theory and applications. Am J Hum Genet 60: 1513-1531.
52. Niu T (2004) Algorithms for inferring haplotypes. Genet Epidemiol 27: 334-347.
53. Devlin B, Risch N (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29: 311-322.
54. Everitt BS (2000) The Analysis of Contingency Tables. Boca Raton: CRC Press.
55. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005) A fine-scale map of recombination rates and hotspots across the human genome. Science 310: 321-324.
56. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29: 229-232.
57. Subrahmanyan L, Eberle MA, Clark AG, Kruglyak L, Nickerson DA (2001) Sequence variation and linkage disequilibrium in the human T-cell receptor beta (TCRB) locus. Am J Hum Genet 69: 381-395.
58. Wang N, Akey JM, Zhang K, Chakraborty R, Jin L (2002) Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. Am J Hum Genet 71: 1227-1234.
59. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, et al. (2002) The structure of haplotype blocks in the human genome. Science 296: 2225-2229.
60. Hudson RR, Kaplan NL (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111: 147-164.
61. Berloff N, Perola M, Lange K (2002) Spline methods for the comparison of physical and genetic maps. J Comput Biol 9: 465-475.
62. Dempster A, Laird N, Rubin D (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society Series B (Methodological) 39: 1-38.
63. Barret J, Fry B, Maller J, Daly M (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics.
64. Sasieni PD (1997) From genotypes to genes: doubling the sample size. Biometrics 53: 1253-1261.
65. Guedj M, Nuel G, Prum B (2008) A note on allelic tests in case-control association studies. Ann Hum Genet 72: 407-409.
66. Lettre G, Lange C, Hirschhorn JN (2007) Genetic model testing and statistical power in population-based association studies of quantitative traits. Genet Epidemiol 31: 358-362.
67. Modin H, Olsson W, Hillert J, Masterman T (2004) Modes of action of HLA-DR susceptibility specificities in multiple sclerosis. Am J Hum Genet 74: 1321-1322.
68. Barcellos LF, Oksenberg JR, Begovich AB, Martin ER, Schmidt S, et al. (2003) HLA-DR2 dose effect on susceptibility to multiple sclerosis and influence on disease course. Am J Hum Genet 72: 710-716.
69. González JR, Armengol L, Guinó E, Solé X, Moreno V (2008) SNPassoc: SNPs-based whole genome association studies.
70. Team RDC (2007) R: A Language and Environment for Statistical Computing. Vienna, Australia: R Foundation for Statistical Computing,.
71. Ahlbom A, Alfredsson L (2005) Interaction: A word with two meanings creates confusion. Eur J Epidemiol 20: 563-564.
72. Pagano M, Gauvreau K (2000) Principles of biostatistics: Brooks/Cole. 525 p.

73. Hutchison KE, Stallings M, McGeary J, Bryan A (2004) Population stratification in the candidate gene study: fatal threat or red herring? Psychol Bull 130: 66-79.
74. Clayton D (2003) Population Association. In: Balding D, Bishop M, Cannings C, editors. Handbook of Statistical Genetics. West Sussex: John Wiley & Sons, Ltd. pp. 939-960.
75. Lehmann EL (2006) Nonparametrics: Statistical Methods Based on Ranks Springer
76. Cox DR (1972) Regression Models and Life Tables. Journal of the Royal Statistical Society Series B: 187–220.
77. Bertrams J, Kuwert E (1972) HL-A antigen frequencies in multiple sclerosis. Significant increase of HL-A3, HL-A10 and W5, and decrease of HL-A12. Eur Neurol 7: 74-78.
78. Naito S, Namerow N, Mickey MR, Terasaki PI (1972) Multiple sclerosis: association with HL-A3. Tissue Antigens 2: 1-4.
79. Dyment DA, Herrera BM, Cader MZ, Willer CJ, Lincoln MR, et al. (2005) Complex interactions among MHC haplotypes in multiple sclerosis: susceptibility and resistance. Hum Mol Genet 14: 2019-2026.
80. Barcellos LF, Sawcer S, Ramsay PP, Baranzini SE, Thomson G, et al. (2006) Heterogeneity at the HLA-DRB1 locus and risk for multiple sclerosis. Hum Mol Genet 15: 2813-2824.
81. Ramagopalan SV, Morris AP, Dyment DA, Herrera BM, DeLuca GC, et al. (2007) The inheritance of resistance alleles in multiple sclerosis. PLoS Genet 3: 1607-1613.
82. Yeo TW, De Jager PL, Gregory SG, Barcellos LF, Walton A, et al. (2007) A second major histocompatibility complex susceptibility locus for multiple sclerosis. Ann Neurol.
83. Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, et al. (2004) Gene map of the extended human MHC. Nat Rev Genet 5: 889-899.
84. Janeway CA, Travers P, Walport M, Shlomchik M (2001) Immunobiology: Garland Publishing.
85. Weinzierl AO, Rudolf D, Hillen N, Tenzer S, van Endert P, et al. (2008) Features of TAP-independent MHC class I ligands revealed by quantitative mass spectrometry. Eur J Immunol 38: 1503-1510.
86. Henderson RA, Michel H, Sakaguchi K, Shabanowitz J, Appella E, et al. (1992) HLA-A2.1-associated peptides from a mutant cell line: a second pathway of antigen presentation. Science 255: 1264-1266.
87. Boname JM, May JS, Stevenson PG (2005) The murine gamma-herpesvirus-68 MK3 protein causes TAP degradation independent of MHC class I heavy chain degradation. Eur J Immunol 35: 171-179.
88. Friese MA, Jakobsen KB, Friis L, Etzensperger R, Craner MJ, et al. (2008) Opposing effects of HLA class I molecules in tuning autoreactive CD8+ T cells in multiple sclerosis. Nat Med 14: 1227-1235.
89. Honma K, Parker KC, Becker KG, McFarland HF, Coligan JE, et al. (1997) Identification of an epitope derived from human proteolipid protein that can induce autoreactive CD8+ cytotoxic T lymphocytes restricted by HLA-A3: evidence for cross-reactivity with an environmental microorganism. J Neuroimmunol 73: 7-14.
90. Masterman T, Ligers A, Olsson T, Andersson M, Olerup O, et al. (2000) HLA-DR15 is associated with lower age at onset in multiple sclerosis. Ann Neurol 48: 211-219.
91. Celius EG, Harbo HF, Egeland T, Vartdal F, Vandvik B, et al. (2000) Sex and age at diagnosis are correlated with the HLA-DR2, DQ6 haplotype in multiple sclerosis. J Neurol Sci 178: 132-135.
92. Hensiek AE, Sawcer SJ, Feakes R, Deans J, Mander A, et al. (2002) HLA-DR 15 is associated with female sex and younger age at diagnosis in multiple sclerosis. J Neurol Neurosurg Psychiatry 72: 184-187.
93. Weatherby SJ, Thomson W, Pepper L, Donn R, Worthington J, et al. (2001) HLA-DRB1 and disease outcome in multiple sclerosis. J Neurol 248: 304-310.
94. Olerup O, Hillert J, Fredrikson S, Olsson T, Kam-Hansen S, et al. (1989) Primarily chronic progressive and relapsing/remitting multiple sclerosis: two immunogenetically distinct disease entities. Proc Natl Acad Sci U S A 86: 7113-7117.

95. de la Concha EG, Arroyo R, Crusius JB, Campillo JA, Martin C, et al. (1997) Combined effect of HLA-DRB1*1501 and interleukin-1 receptor antagonist gene allele 2 in susceptibility to relapsing/remitting multiple sclerosis. J Neuroimmunol 80: 172-178.
96. Weinshenker BG, Santrach P, Bissonet AS, McDonnell SK, Schaid D, et al. (1998) Major histocompatibility complex class II alleles and the course and outcome of MS: a population-based study. Neurology 51: 742-747.
97. IMSGC (2008) Refining genetic associations in multiple sclerosis. Lancet Neurol 7: 567-569.
98. Zoledziewska M, Costa G, Pitzalis M, Cocco E, Melis C, et al. (2008) Variation within the CLEC16A gene shows consistent disease association with both multiple sclerosis and type 1 diabetes in Sardinia. Genes Immun.
99. Kleine TO, Benes L (2006) Immune surveillance of the human central nervous system (CNS): different migration pathways of immune cells through the blood-brain barrier and blood-cerebrospinal fluid barrier in healthy persons. Cytometry A 69: 147-151.
100. Ye SK, Maki K, Kitamura T, Sunaga S, Akashi K, et al. (1999) Induction of germline transcription in the TCRgamma locus by Stat5: implications for accessibility control by the IL-7 receptor. Immunity 11: 213-223.
101. Morozova O, Marra MA (2008) Applications of next-generation sequencing technologies in functional genomics. Genomics 92: 255-264.
102. Affymetrix I (2007) GeneChip® Human Genome U133 Arrays: The Most Comprehensive Coverage of the Human Genome in Two
Flexible Formats: Single-array Cartridges and Multi-array Plates. Santa Clara: Affymetrix. pp. GeneChip® Human Genome U133 Arrays - Data Sheet.
103. Affymetrix I (2006) GeneChip® Exon Array System for Human, Mouse, and Rat: Genome-Wide Gene Expression and Alternative Splicing Profiling on a Single Array. Santa Clara: Affymetrix. pp. GeneChip® Exon Array System for Human, Mouse, and Rat - Data sheet.
104. Affymetrix I GeneChip® Human Tiling Arrays: Highest resolution tiling arrays for most accurate mapping of protein/DNA interactions and novel transcript discovery. Santa Clara: Affymetrix. pp. GeneChip® Human Tiling Arrays - Data Sheet.
105. Affymetrix I (2003) Design and Performance of the GeneChip® Human Genome U133 Plus 2.0 and Human Genome U133A 2.0 Arrays. Affymetrix Inc. pp. Human Genome U133 Plus 132.130 technical notes.
106. Raj A, van Oudenaarden A (2008) Nature, nurture, or chance: stochastic gene expression and its consequences. Cell 135: 216-226.
107. Thompson KL, Pine PS, Rosenzweig BA, Turpaz Y, Retief J (2007) Characterization of the effect of sample quality on high density oligonucleotide microarray data using progressively degraded rat liver RNA. BMC Biotechnol 7: 57.
108. Cepok S, Jacobsen M, Schock S, Omer B, Jaekel S, et al. (2001) Patterns of cerebrospinal fluid pathology correlate with disease progression in multiple sclerosis. Brain 124: 2169-2176.
109. Vercellino M, Votta B, Condello C, Piacentino C, Romagnolo A, et al. (2008) Involvement of the choroid plexus in multiple sclerosis autoimmune inflammation: a neuropathological study. J Neuroimmunol 199: 133-141.
110. Hafler DA, Weiner HL (1987) T cells in multiple sclerosis and inflammatory central nervous system diseases. Immunol Rev 100: 307-332.
111. Winges KM, Gilden DH, Bennett JL, Yu X, Ritchie AM, et al. (2007) Analysis of multiple sclerosis cerebrospinal fluid reveals a continuum of clonally related antibody-secreting cells that are predominantly plasma blasts. J Neuroimmunol 192: 226-234.
112. Corcione A, Aloisi F, Serafini B, Capello E, Mancardi GL, et al. (2005) B-cell differentiation in the CNS of patients with multiple sclerosis. Autoimmun Rev 4: 549-554.
113. Jacobsen M, Cepok S, Quak E, Happel M, Gaber R, et al. (2002) Oligoclonal expansion of memory CD8+ T cells in cerebrospinal fluid from multiple sclerosis patients. Brain 125: 538-550.

114. Achiron A, Gurevich M (2006) Peripheral blood gene expression signature mirrors central nervous system disease: the model of multiple sclerosis. Autoimmun Rev 5: 517-522.
115. Sarkijarvi S, Kuusisto H, Paalavuo R, Levula M, Airla N, et al. (2006) Gene expression profiles in Finnish twins with multiple sclerosis. BMC Med Genet 7: 11.
116. Avasarala JR, Chittur SV, George AD, Tine JA (2008) Microarray analysis in B cells among siblings with/without MS - role for transcription factor TCF2. BMC Med Genomics 1: 2.
117. Arthur AT, Armati PJ, Bye C, Consortium SM, Heard RN, et al. (2008) Genes implicated in multiple sclerosis pathogenesis from consilience of genotyping and expression profiles in relapse and remission. BMC Med Genet 9: 17.
118. Satoh J, Misawa T, Tabunoki H, Yamamura T (2008) Molecular network analysis of T-cell transcriptome suggests aberrant regulation of gene expression by NF-kappaB as a biomarker for relapse of multiple sclerosis. Dis Markers 25: 27-35.
119. Achiron A, Feldman A, Mandel M, Gurevich M (2007) Impaired expression of peripheral blood apoptotic-related gene transcripts in acute multiple sclerosis relapse. Ann N Y Acad Sci 1107: 155-167.
120. Malmestrom C, Lycke J, Haghighi S, Andersen O, Carlsson L, et al. (2008) Relapses in multiple sclerosis are associated with increased CD8+ T-cell mediated cytotoxicity in CSF. J Neuroimmunol 196: 159-165.
121. Bomprezzi R, Ringner M, Kim S, Bittner ML, Khan J, et al. (2003) Gene expression profile in multiple sclerosis patients and healthy controls: identifying pathways relevant to disease. Hum Mol Genet 12: 2191-2199.
122. Mandel M, Gurevich M, Pauzner R, Kaminski N, Achiron A (2004) Autoimmunity gene expression portrait: specific signature that intersects or differentiates between multiple sclerosis and systemic lupus erythematosus. Clin Exp Immunol 138: 164-170.
123. Annibali V, Di Giovanni S, Cannoni S, Giugni E, Bomprezzi R, et al. (2007) Gene expression profiles reveal homeostatic dynamics during interferon-beta therapy in multiple sclerosis. Autoimmunity 40: 16-22.
124. Sellebjerg F, Datta P, Larsen J, Rieneck K, Alsing I, et al. (2008) Gene expression analysis of interferon-beta treatment in multiple sclerosis. Mult Scler 14: 615-621.
125. van Baarsen LG, Vosslamber S, Tijssen M, Baggen JM, van der Voort LF, et al. (2008) Pharmacogenomics of interferon-beta therapy in multiple sclerosis: baseline IFN signature determines pharmacological differences between patients. PLoS ONE 3: e1927.
126. Rani MR, Shrock J, Appachi S, Rudick RA, Williams BR, et al. (2007) Novel interferon-beta-induced gene expression in peripheral blood cells. J Leukoc Biol 82: 1353-1360.
127. Singh MK, Scott TF, LaFramboise WA, Hu FZ, Post JC, et al. (2007) Gene expression changes in peripheral blood mononuclear cells from multiple sclerosis patients undergoing beta-interferon therapy. J Neurol Sci 258: 52-59.
128. Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, et al. (2006) The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol Biol 7: 3.
129. Irizarry RA, Wu Z, Jaffee HA (2006) Comparison of Affymetrix GeneChip expression measures. Bioinformatics 22: 789-794.
130. Tan PK, Downey TJ, Spitznagel EL, Jr., Xu P, Fu D, et al. (2003) Evaluation of gene expression measurements from commercial microarray platforms. Nucleic Acids Res 31: 5676-5684.
131. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4: 249-264.
132. Millenaar FF, Okyere J, May ST, van Zanten M, Voesenek LA, et al. (2006) How to decide? Different methods of calculating gene expression from short oligonucleotide array data will give different results. BMC Bioinformatics 7: 137.
133. Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F (2004) A Model-Based Background Adjustment for Oligonucleotide Expression Arrays Journal of the American Statistical Association 99: 909-917.

134. Cambon AC, Khalyfa A, Cooper NG, Thompson CM (2007) Analysis of probe level patterns in Affymetrix microarray data. BMC Bioinformatics 8: 146.
135. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19: 185-193.
136. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, et al. (2003) Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 31: e15.
137. Barash Y, Dehan E, Krupsky M, Franklin W, Geraci M, et al. (2004) Comparative analysis of algorithms for signal quantitation from oligonucleotide microarrays. Bioinformatics 20: 839-846.
138. Seo J, Gordish-Dressman H, Hoffman EP (2006) An interactive power analysis tool for microarray hypothesis testing and generation. Bioinformatics 22: 808-814.
139. Wilson CL, Miller CJ (2005) Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis. Bioinformatics 21: 3683-3685.
140. Huber W, von Heydebreck A, Vingron M (2003) Analysis of Microarray Gene Expression Data. In: Balding D, Bishop M, Cannings C, editors. Handbook of Statistical Genetics. West Sussex: John Wiley & Sons, Ltd. pp. 162-187.
141. Dudoit S, Yang YH, Callow MJ, Speed TP (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica 12: 111-139.
142. Culhane AC, Thioulouse J, Perriere G, Higgins DG (2005) MADE4: an R package for multivariate analysis of gene expression data. Bioinformatics 21: 2789-2790.
143. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 98: 5116-5121.
144. Smyth GK (2005) Limma: linear models for microarray data. In: Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W, editors. Bioinformatics and Computational Biology Solutions using R and Bioconductor. New York: Springer. pp. 397-420.
145. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3: Article3.
146. Breitling R, Armengaud P, Amtmann A, Herzyk P (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett 573: 83-92.
147. Parodi S, Muselli M, Fontana V, Bonassi S (2003) ROC curves are a suitable and flexible tool for the analysis of gene expression profiles. Cytogenet Genome Res 101: 90-91.
148. Benjamini Y, Hochberg Y (1995) Controlling the false disovery rate: a practical and powerful approach to multiple testing. J R Statist Soc B 57: 298-300.
149. Pollard KS GY, Taylor S, Dudoit S multtest: Resampling-based multiple hypothesis testing. .
150. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29.
151. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27-30.
152. Wingender E, Chen X, Hehl R, Karas H, Liebich I, et al. (2000) TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res 28: 316-319.
153. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, et al. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 34: D108-110.
154. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, et al. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 34: 267-273.
155. Gentleman RC, Falcon, S. (2007) Category: Category Analysis.
156. Carlson M, Falcon S, Pages H, Li N hgu133plus2.db: Affymetrix Human Genome U133 Plus 2.0 Array annotation data (chip hgu133plus2).
157. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, et al. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 21: 3439-3440.

158. Gautier L, Moller M, Friis-Hansen L, Knudsen S (2004) Alternative mapping of probes to genes for Affymetrix chips. BMC Bioinformatics 5: 111.
159. Ferrari F, Bortoluzzi S, Coppe A, Sirota A, Safran M, et al. (2007) Novel definition files for human GeneChips based on GeneAnnot. BMC Bioinformatics 8: 446.
160. Gibson UE, Heid CA, Williams PM (1996) A novel method for real time quantitative RT-PCR. Genome Res 6: 995-1001.
161. Heid CA, Stevens J, Livak KJ, Williams PM (1996) Real time quantitative PCR. Genome Res 6: 986-994.
162. Pfaffl MW (2001) A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res 29: e45.
163. Pfaffl MW, Horgan GW, Dempfle L (2002) Relative expression software tool (REST) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. Nucleic Acids Res 30: e36.
164. Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 25: 402-408.
165. Morrison TB, Weis JJ, Wittwer CT (1998) Quantification of low-copy transcripts by continuous SYBR Green I monitoring during amplification. Biotechniques 24: 954-958, 960, 962.
166. Morey JS, Ryan JC, Van Dolah FM (2006) Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR. Biol Proced Online 8: 175-193.
167. Pfaffl MW (2006) Relative Quantification. In: Dorak T, editor. Real time PCR: Taylor & Francis Group. pp. 63-82.
168. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, et al. (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol 3: RESEARCH0034.
169. Wang L, Blasic JR, Jr., Holden MJ, Pires R (2005) Sensitivity comparison of real-time PCR probe designs on a model DNA plasmid. Anal Biochem 344: 257-265.
170. Verstraeten VL, Broers JL, Ramaekers FC, van Steensel MA (2007) The nuclear envelope, a key structure in cellular integrity and gene expression. Curr Med Chem 14: 1231-1248.
171. Tham E, Gielen AW, Khademi M, Martin C, Piehl F (2006) Decreased expression of VEGF-A in rat experimental autoimmune encephalomyelitis and in cerebrospinal fluid mononuclear cells from patients with multiple sclerosis. Scand J Immunol 64: 609-622.
172. Koning N, Bo L, Hoek RM, Huitinga I (2007) Downregulation of macrophage inhibitory molecules in multiple sclerosis lesions. Ann Neurol 62: 504-514.
173. Whitney LW, Ludwin SK, McFarland HF, Biddison WE (2001) Microarray analysis of gene expression in multiple sclerosis and EAE identifies 5-lipoxygenase as a component of inflammatory lesions. J Neuroimmunol 121: 40-48.
174. Baranzini SE, Elfstrom C, Chang SY, Butunoi C, Murray R, et al. (2000) Transcriptional analysis of multiple sclerosis brain lesions reveals a complex pattern of cytokine expression. J Immunol 165: 6576-6582.
175. Gveric D, Hanemaaijer R, Newcombe J, van Lent NA, Sier CF, et al. (2001) Plasminogen activators in multiple sclerosis lesions: implications for the inflammatory response and axonal damage. Brain 124: 1978-1988.
176. Blasi F (1997) uPA, uPAR, PAI-1: key intersection of proteolytic, adhesive and chemotactic highways? Immunol Today 18: 415-417.
177. Danilova N, Sakamoto KM, Lin S (2008) Ribosomal protein S19 deficiency in zebrafish leads to developmental abnormalities and defective erythropoiesis through activation of p53 protein family. Blood.
178. Gazouli M, Kokotas S, Zoumpourlis V, Zacharatos P, Mariatos G, et al. (2002) The complement inhibitor CD59 and the lymphocyte function-associated antigen-3 (LFA-3, CD58) genes possess functional binding sites for the p53 tumor suppressor protein. Anticancer Res 22: 4237-4241.