The Complex Genetics of Multiple Sclerosis
A preliminary study of MS-associated SNPs prior to a larger genotyping project

Simon Söderholm

Examinator, Mika Gustafsson
Tutor, Andreas Tjärnberg
Abstract

Biomedical research have been revolutionized by recent technological advances, both in the fields of molecular biology and computer science, turning the biomolecular and genetic research into “big data science”. One of the main objectives have been to improve our understanding of complex human diseases. Among those diseases, multiple sclerosis (MS) is considered as one of the most common. MS is a chronic autoimmune disease that cause inflammation and damage to the central nervous system. In this study, a set of bioinformatics analyses have been conducted on SNP data, as an initial step to gain more information prior to an upcoming genotyping project. The results showed extensive regulatory properties for the 761 selected SNPs, which is consistent with current scientific knowledge, and also identified another 332 SNPs in linkage to these. However, during the study some issues have also been identified, which need to be addressed going forward.

A short comment on essential abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNS</td>
<td>Central Nervous System</td>
</tr>
<tr>
<td>eQTL</td>
<td>Expression Quantitative Trait Loci</td>
</tr>
<tr>
<td>GWAS</td>
<td>Genome-Wide Association Study</td>
</tr>
<tr>
<td>HLA</td>
<td>Human Leukocyte Antigen</td>
</tr>
<tr>
<td>IGSR</td>
<td>The International Genome Sample Resource</td>
</tr>
<tr>
<td>IKE</td>
<td>Department of Clinical and Experimental Medicine at Linköping University</td>
</tr>
<tr>
<td>IMSGC</td>
<td>International Multiple Sclerosis Genetics Consortium</td>
</tr>
<tr>
<td>LD</td>
<td>Linkage Disequilibrium</td>
</tr>
<tr>
<td>MHC</td>
<td>Major Histocompatibility Complex</td>
</tr>
<tr>
<td>MS</td>
<td>Multiple Sclerosis</td>
</tr>
<tr>
<td>NGS</td>
<td>Next-Generation Sequencing</td>
</tr>
<tr>
<td>PPMS</td>
<td>Primary Progressive Multiple Sclerosis</td>
</tr>
<tr>
<td>RRMS</td>
<td>Relapsing-Remitting Multiple Sclerosis</td>
</tr>
<tr>
<td>SNP</td>
<td>Single Nucleotide Polymorphism</td>
</tr>
<tr>
<td>SPMS</td>
<td>Secondary Progressive Multiple Sclerosis</td>
</tr>
<tr>
<td>MAF</td>
<td>Mean Allele Frequency</td>
</tr>
</tbody>
</table>
Content
1. Introduction .................................................................................................................. 4
2. Theoretical background ............................................................................................... 5
   2.1 Multiple sclerosis .................................................................................................... 5
   2.2 MS Pathology ......................................................................................................... 6
   2.3 MS genetics ............................................................................................................. 6
     2.3.1 MHC and the HLA system ................................................................................ 7
     2.3.2 Linkage disequilibrium (LD) and its significance for HLA inheritance .......... 9
   2.4 The role of T cells and T cell differentiation .......................................................... 10
   2.5 “Missing heritability” – epigenetics and environmental influence ......................... 11
3. Materials and methods ............................................................................................... 12
   3.1 Brief comment on the SNEMS project methods ....................................................... 12
   3.2 Preliminary study methods ..................................................................................... 13
     3.2.1 Selection of significant SNPs ............................................................................. 13
     3.2.2 Validation of linkage disequilibrium for the SNPs ........................................... 13
     3.2.3 Mapping SNPs to chromosomal positions ......................................................... 14
     3.2.4 Examination of regulatory properties ............................................................... 14
     3.2.5 Allele frequencies and gene identification ......................................................... 14
     3.2.6 Mapping genes to biological pathways ............................................................. 15
4. Results .......................................................................................................................... 15
   4.1 Examination of linkage disequilibrium .................................................................... 16
   4.2 Mapping of SNPs to chromosomal positions .......................................................... 16
   4.3 Regulatory scoring of SNPs .................................................................................... 17
   4.4 Allele frequencies .................................................................................................. 17
   4.4 Examination of identified SNP-containing genes .................................................... 17
   4.5 Mapping genes to biological pathways .................................................................. 19
5. Discussion, conclusions and ideas for the planned project ........................................... 21
6. References .................................................................................................................... 25
   6.1 Literature ............................................................................................................... 25
   6.2 R programming packages ....................................................................................... 29
   6.3 Databases ................................................................................................................ 29
1. Introduction

In 2003 the sequencing of the entire human genome was completed through the Human Genome Project. Since then, the biomedical research have been revolutionized by new technological advances, such as high-throughput measurement of gene expression and detection of genetic variation, next-generation sequencing and genome-wide associations studies (Hofker et al, 2014). All these technological innovations has provided the tools for genetic and biomolecular research to generate large amounts of data and becoming “big data science”. This in turn has led to the development of the still relatively new research field of system biology, which studies the structure and dynamics of biological systems (Kitano, 2002).

The new knowledge obtained from this development has made it possible to begin to understand the genetic and biochemical background of human diseases. One of the major aims of the genetic research is to use the abundance of biological information to expand the understanding of common complex diseases. Complex diseases, which include cancer, Alzheimer’s disease, autoimmune diseases, type 2 diabetes and cardiovascular diseases, are caused by combinations of and interactions between many factors, both genetic and environmental (Hofker et al, 2014). These disorders constitute a significant burden on healthcare and society, and medical treatment are still largely ineffective. Therefore, the goal of the research is to contribute to the development of personalized medicine, where patients’ personal genetic and environmental factors are taken into account in the assessment of disease susceptibility and choice of appropriate treatment.

Multiple sclerosis (MS) is one of the most common complex diseases, whose etiology and symptoms are both variable and unpredictable. MS is an autoimmune disease characterized by inflammation and neurological damage in the central nervous system.

In this thesis, a preliminary study have been conducted prior to an upcoming research project on MS genetics. The aim of this study is to conduct initial analysis of a data set of SNPs, in order to outline their association with MS, as well as to gain information of value for the planned project. This includes genomic mapping of SNPs of interest, investigation of linkage disequilibrium, and identification of known biological pathways associated with these SNPs. Furthermore, the aim is to compile current scientific knowledge of relevance, as well as to conduct a first examination of appropriate methods, tools and procedures prior to the upcoming research project.
2. Theoretical background

In the following section the theoretical foundation for the study is outlined. Here the characteristics of multiple sclerosis is described, as well as current knowledge about its underlying genetic and epigenetic pathogenic associations.

2.1 Multiple sclerosis

Multiple sclerosis (MS) has a history that spans more than a century back in time. Indeed, there are descriptions of people that can be identified as suffering from MS already in medieval literature, but it was when physicians began to study the disease in more scientific terms that MS received a more detailed description. As early as the year 1838, there are illustrations that clearly shows what is today known as MS (Loren & Rolak, 2016).

MS is a chronic neurologic disease with inflammatory and autoimmune characteristics, caused by an uncontrolled immune response. It affects the central nervous system (CNS), i.e. the brain (encephalon) and the spinal cord (medulla spinale), by attacking the axons and destroying their layers of myelin, and to various extend also leads to the degeneration of the axons themselves (Compston & Coles, 2008; Simons et al., 2014; Lassman, 2014; Al-Tubaikh, 2010).

At autopsy of the CNS, hard rubbery areas with a slight pink or grey colour tone are revealed in the white matter (Goldberg, 2012). These lesions have been caused by the loss of myelin and oligodendrocytes and the accumulation of inflammatory cells such as macrophages and lymphocytes (Al-Tubaikh, 2010).

The disease progression of MS may have varied appearances and unpredictable behaviour from case to case. In the vast majority of patients suffering from MS the disease usually begins with recurring episodes of neurological defects, which are separated by symptom-free periods (Goldenberg, 2012). Generally, however, there is also a gradual deterioration of the patient’s condition. Typical symptoms of MS are muscle weakness in arms and legs, local numbness, impaired vision including effects such as double vision and partial or complete loss of vision, and memory and fatigue problems. Other symptoms are bladder dysfunction, which occurs in 90% of disease cases, constipation that affects about 30% of patients, and substantial fatigue, which occurs in 90% of cases and is one of the most common MS-related causes of impaired work capacity (Goldenberg, 2012; Al-Tubaikh, 2010). During the latter part of the course of the disease motoric difficulties and paralysis may develop, as well as slow speech and trouble chewing and swallowing (Sweden Medical Product Agency, 2016-04-01).

Assembling the global prevalence of MS from epidemiological data have long been a complicated task that involves a number of problems. A compilation study conducted by Pugliatti et al. (2002) showed that the highest prevalence of MS is found in Scandinavia and the UK. It has been generally believed for many years that the spread of the disease appears to follow a latitude dependent pattern, but this theory has also been challenged by some scientists. However, in a comprehensive review of MS prevalence made by Simpson et al. (2011) it is suggested that MS prevalence indeed follows a latitude pattern. Simpson’s meta-analysis, which included 650 prevalence surveys from 321 peer-reviewed studies in 59 countries between 1923 and 2009, showed a statistically significant positive association between MS prevalence and latitude. This gradient applied mainly to regions of predominantly European descent, while not discovered in non-European descent regions. This result seems likely, since it already have
been determined that high-risk alleles for MS occurs at higher rates in European populations (Schmidt et al., 2007).

There is currently no cure for MS. Instead, treatments and therapeutic agents aims to improve and maintain the patient’s quality of life (Goldberg, 2012).

2.2 MS Pathology
The last decades have seen extensive efforts of research on the etiology and pathogenesis of MS, but much central information is still not fully understood. This is due to the fact that the availability of appropriate tissue samples from MS patients, where active demyelination and neurodegeneration could be detected, have been limited (Lassmann, 2014). Thus, much of the knowledge regarding the mechanisms of cerebral inflammation and immune-mediated tissue damage have been acquired through analysis of data from experimental models. However, the research of the recent century have provided us with valuable knowledge of MS pathology processes, which has been enhanced by the advances in modern neurobiology and immunology.

In a majority of MS patients, the disease initially follows a relapsing-remitting course (RRMS), characterized by episodic neurological disorder separated by symptom-free periods. After a number of years the disease condition enters a secondary progressive phase (SPMS). Some patients also suffer from a condition called primary progressive MS (PPMS), where the initial relapsing-remitting stage is absent and the disease starts with a constant progression from the very beginning (Mahad et al., 2015).

Studies over the years have suggested that MS is caused by autoimmune factors, where T lymphocytes leaves the immune system and enters the CNS in the beginning phase of lesion formation. T helper cell differentiation have been indicated as having a particular role in the multiple sclerosis pathogenesis (Sawcer et al., 2012). However, there have been several disagreeing ideas for MS progression and disease mechanisms proposed in various reliable studies, and to date no autoimmune reaction specific for MS has been identified (Mahad et al., 2015). Mahad et al. aims in their study to create a consolidated picture, by presenting a model for immunological and neurodegenerative events which contributes to neuronal damage in association with progressive MS.

Axons possess their own production of mitochondria and ATP (Mahad et al., 2015). Many of these mitochondria are assembled at specific sites and is considered to account for the majority of ATP production in the axons. However, the axons are very sensitive to mitochondrial dysfunction, which is a feature of multiple sclerosis demyelination. This leads to decreased ATP production, increased levels of reactive oxygen species, an altered distribution of ions and ultimately resulting in cell degeneration. Although the most typical form of brain injury is demyelination of axons, the major pathological consequence of progressive MS is atrophy of the brain.

2.3 MS genetics
The technological development in recent decades have revolutionized the field of genetic research. The use of genome-wide association studies (GWAS) and the advent of next generation sequencing (NGE) has resulted in a significant improvement of our understanding of genetics and molecular biological processes (Gibson, 2010; Meldrum et al, 2011). Furthermore, the rapid development of analysis tools within the field of bioinformatics has also had a major impact on the possibilities for advanced genomic research and the study of complex diseases. In the case of multiple sclerosis, a total of 15 GWASs have been conducted to date,
of which 14 have been mentioned in previous articles (Sawcer et al, 2014; Didonna et al, 2015) and one presented in a recent published study (Giacalone et al, 2015).

However, there are some shortcomings of GWAS that also have been mentioned. For example, Gibson (2010) points out that although genome-wide association studies have contributed to the identification of genetic variants for a large number of complex diseases, these types of studies have only been able to explain just a fraction of the disease heritability and it is likely to assume that genetic prediction alone will not be sufficiently clinically informative. Nevertheless, GWAS has led to significant advancements and revealed important part of the genetic basis of many complex diseases, including MS.

One of the more extensive GWAS, conducted by The International Multiple Sclerosis Genetics Consortium (IMSGC), was published in 2011 and involved 9772 cases collected in 15 different countries by 23 research groups (Sawcer et al, 2012). In this comprehensive study a total of over 440 000 SNPs where analysed, of which 102 were identified as significantly associated with MS. These results form the basis for the study in this thesis, as well as for the upcoming genotyping project.

Research has focused on the study of disease-related single nucleotide polymorphisms (SNPs), and this work have been enhanced by recent year's development of bio-statistical methods and high-throughput genotyping. Through internationally-driven research projects it has been possible to identify more than 100 common variants with significant association with multiple sclerosis. Studies have shown that these genetic loci to a large extent are found in regulatory regions and that they affect genes of importance for immunological processes. In addition to association with MS, these common variants has also been shown to have association with other autoimmune diseases (Sawcer et al, 2014).

Genomic regions of particular importance for the occurrence of MS is the major histocompatibility complex (MHC), in humans also known as the human leukocyte antigen (HLA) region.

2.3.1 MHC and the HLA system

The Major Histocompatibility Complex (MHC), illustrated in figure 1 below, is a genomic region with high gene density spanning approximately 3600 base pairs and is located on the short arm of chromosome 6 (6p21.3) (Didonna et al, 2015; Alcock et al, 2002; Choo, 2007). MHC is considered as the most important genetic region in regards to many diseases common in humans, such as autoimmune disorders, infection and inflammation. Numerous studies over the years has identified variations in this region as the single major risk factor for MS susceptibility (Alcock et al, 2002; Horton et al, 2008; Sawcer et al, 2012).
However, the MHC also constitute the most variable region of the human genome, with a high level of gene clustering, polymorphisms and linkage disequilibrium (ie. Correlation between linked variants), which always has brought complications in studies of association between disease phenotypes and MHC (Allcock et al, 2002).

For this reason, the MHC Haplotype Project was assembled, and conducted between 2000 and 2006 at the Sanger Institute. The aim of that project was to determine the entire DNA sequence of a number of common MHC haplotypes, as well as to identify all SNPs of medical importance. The consortium analysed eight haplotypes relevant for type 1 diabetes and multiple sclerosis, in order to present a comprehensive and adequate annotated reference MHC haplotype against which variations in the other seven could be related (Allcock et al, 2002; Horton et al, 2008). These are considered as the most common MHC haplotypes in the European population and are designated as PGE (the reference haplotype), COX, APD, AZH, MANN, SSTO, QBL and MCF.

Human leukocyte antigens (HLAs) is a form of antigen-presenting protein molecules on the cell surface of all nucleus-containing cells in the human body. Their task is to present peptides to immune cells which upon detection of a foreign antigen bound to HLAs launches an immune response. The HLA system represents the human-exclusive gene complex occupying the MHC region, and whose encoded proteins play an important role in the regulation of immune response process (Choo, 2007).

The human MHC region is divided into three sub-regions, which together comprises 224 identified loci. Of those, about 150 have been estimated as expressed and 40% encodes proteins with immunological functions (Choo, 2007; Didonna & Oksenberg, 2015).

The class I sub-region contains six genes, A – G, and eleven pseudogenes (i.e. unexpressed genes), H – Y. Of those, the HLA-A, HLA-B and HLA-C genes have been associated with MS. HLAs corresponding to this class II sub-region present peptides from inside the cell originated from processes such as virus infections. These molecules are expressed on the surface of most cells that possess a nucleus.

However, of most interest regarding MS susceptibility is the class II sub-region, containing genes (DP, DM, DOA, DOB, DQ and DR) for the classical HLA II molecules which plays an important role in the identification and presentation of antigens from outside the cell to T cells (Schmidt et al, 2007; Robinson et al, 2015), and studies have proven these molecules to be associated with a number of immunological related diseases. Of those, the three gene families DP, DQ and DR, each representing a type of HLA class II molecules, are of certain relevance for MS disease. Such class II molecules are expressed only on antigen-presenting cells, B-cells and T-cells (Choo, 2007). Among all the identified genes of the MHC region the strongest MS association have been found for the HLA-DRB1 gene, specifically the DRB1*15:01 allele, part

![Figure 1: Illustration of the human chromosome 6, indicating the MHC region and the HLA genes (created by Philip Deitiker, released into the public domain).](image-url)
of the DR15 haplotype (DRB1*1501-DQA1*0102-DQB1*0601) which have been linked to a number of diseases and medical conditions (Sawcer et al, 2012; Hollenbach & Oksenberg, 2015; Schmidt et al, 2007).

The class III sub-region of HLA does not produce any HLAs. Instead, it contains components of importance for the complement system, which is a part of the immune system that helps to improve the antibacterial activity of antibodies (Choo 2007; Janeway et al, 2001).

As earlier mentioned, the HLA region constitute a part of the DNA with high genetic diversity, which is believed to result in beneficial effects on the immune system, by providing resistance to a broader spectrum of pathogenic antigens (Schmidt et al, 2007). According to the IMGT/HLA database a total of 14 473 HLA alleles have been identified to date, 10 730 in the class I sub-region and 3 743 in the class II sub-region (update 2016-04-15).

2.3.2 Linkage disequilibrium (LD) and its significance for HLA inheritance

The term linkage disequilibrium (LD) refers to the non-random association between alleles located at different loci (Slatkin, 2008). It is a genetic feature of importance for scientific fields such as evolutionary biology and human genetics.

There have been different definitions of linkage disequilibrium, focusing on different aspects of non-random association. However, a general description could be given in the following way. Suppose that in a certain population allele A occurs with frequency $P_A$ at a locus among reproductively produced gametes, while allele B occurs with the frequency $P_B$ at another locus. Then let $P_{AB}$ represent the frequency of A and B occurring together in the same gamete. If the probability to find A and B together in a gamete chosen at random ($P_{AB}$) is the same as the probability that they independently occurs in the same gamete ($P_A P_B$), the association between the two alleles are said to be completely random and there is a linkage equilibrium (LE). However, if $P_{AB}$ differ from $P_A P_B$ in any respect, it means that there is non-random association between allele A and allele B, and that there is linkage disequilibrium between them.

Central to all types of definitions of LD is the quantity $D_{AB}$ which is the coefficient of linkage disequilibrium (Slatkin, 2008) and is calculated by the equation:

$$D_{AB} = P_{AB} - P_A P_B$$  \hspace{1cm} (Equation 1)

Where $P_{AB}$ and $P_A P_B$ are frequencies as described above. However, although the $D$-value constitutes a suitable characterization of non-random association between two alleles, it is less convenient when comparing LD between pairs of loci because the $D$ value adopts different ranges of possible values depending on the individual allele frequencies of each pair. Therefore, $D'$ have been defined, which is the ratio between $D$ and its maximum possible allelic frequency-dependent value. The advantage is that a value between 0 and 1 is obtained, regardless of the allele frequencies.

Additionally, another way to quantify LD is the commonly used $r^2$ value according to:

$$r^2 = \frac{D^2}{P_A(1-P_A)P_B(1-P_B)}$$  \hspace{1cm} (Equation 2)
The $r^2$ value (also known as coefficient of determination) is the square of the sample correlation coefficient which is widely used in science to describe the degree of linear dependence between two variables. The $r^2$ value is fairly easy to understand and is therefore used by many scientists.

SNPs located close to each other in the genome are often also in LD with each other, but it has also been found that SNPs associated with a certain phenotype indeed can be in full LD with SNPs a couple of hundred kilobases away (Schaub et al, 2012).

Considering random combinations of antigens from different HLA loci, it is evident that the number of possible arrangements are enormous in this region. However, genes in the HLA region have shown to be closely linked to each other, and the MHC region is inherited in its entirety through Mendelian inheritance from each parent. Because if this extensive linkage disequilibrium, gene variants tend to be associated with each other, forming a number of common haplotypes. Certain haplotypes occurs more frequently within different populations than could be expected by chance, making linkage disequilibrium an important parameter to take into consideration in the study of this genomic region (Choo, 2007; Schmidt et al, 2007).

For example, in a study conducted by Pröll et al. (2011) a total of 885 576 equally genomic distributed SNPs where analysed and compared through next generation sequencing. It could be observed that 62% of the analysed loci located outside of MHC was shared between donor and recipient, while 86% of those located within the MHC was shared, suggesting a higher linkage disequilibrium of loci found in the MHC region.

2.4 The role of T cells and T cell differentiation

T cells, also known as T lymphocytes, belongs to a subset of white blood cells and holds great significance in the cellular part of the immune system. Unlike other types of lymphocytes the T cells can be identified through the presence of T cell receptors on the surface of the cells (Berg et al, 2012; Pierce, 2014).

These cells are present in a number of different types. These include T helper cells ($T_H$ cells), which have an assisting function in immunological processes, such as activation of cytotoxic T cells and macrophages. A characteristic feature of these cells is the expression of the CD4 glycoprotein on their cell surfaces, and therefore $T_H$ cells are also called CD4$^+$ cells.

Another important form of T cells is the cytotoxic T cells ($T_C$ cells), which are responsible for the search and destruction of tumour cells and cells infected by viruses. Cytotoxic cells are also known as CD8$^+$ cells because, like the $T_H$ cells, they also express a specific cell surface protein, in this case the CD8 glycoprotein.

In the human blood a majority of the T cells are so-called naive T cells, which are T cells that have not yet been activated by any antigen. After activation the T cells are transformed to their various types (such as $T_H1$, $T_H2$ etc.), and this T cell differentiation has been of interest in the research of autoimmune diseases. $T_H$ cells are activated by antigens presented by MHC class II molecules, and $T_C$ cells are activated when presented antigens by MHC class I molecules.

In the case of multiple sclerosis, regulatory T cells ($T_{REG}$ cells) have been identified as a key operator in the autoimmunological processes of CNS inflammation, demyelination and axonal damage (Zozulya & Weindl, 2008). The role of $T_{REG}$ cells is to prevent harmful immune responses to occur, both directed against foreign and the body’s own antigens, by controlling the number and behaviour of autoreactive T cells (Mills, 2004). The MS disease are believed
to arise as a result of a disturbed balance between tissue damaging cells and controlling T\textsubscript{REG} cells (Zozulya & Weindl, 2008).

In the comprehensive GWAS conducted by IMSGC it was found that genes related to lymphocyte function and T cell activation and production had the most significant enrichment nearest the association region SNPs, linking T helper cell differentiation to the MS pathogenesis (Sawcer et al, 2012).

In a study conducted by Arneth et al. (2015) blood samples from MS patients and healthy subjects were analysed for activity of CD4\textsuperscript{+} and CD8\textsuperscript{+} T cells by injection of human myelin basic protein (MBP). The study showed that T cells from MS patients had a significantly stronger response to MBP than those from healthy persons, concluding that early activation of CD4\textsuperscript{+} and CD8\textsuperscript{+} T cells by MBP constitute a major auto-antigen response process in MS disease.

2.5 “Missing heritability” – epigenetics and environmental influence

An interesting fact is that although genome-wide association studies based on comparison of SNP allele frequencies has become a common and powerful method for the study of MS genetics, the so far identified loci of association have only been able to explain a fraction of the heritability (Sawcer et al, 2014; Bashinskaya et al, 2015). Recently conducted sibling studies have suggested that 27% of the total MS heritability is accounted for by the MS risk loci identified to date, and that the HLA-DRB1*15:01 allele alone accounts for 20% of this heritability (Lill, 2014). The lack of genetic basis for heritability have led to the proposition of a number of possible explanations. Inadequate GWAS design, unidentified interactions between risk factors and undiscovered risk alleles are some explanations that have been suggested (Bashinskaya et al, 2015; Sawcer et al, 2014).

This “missing heritability” have been particularly evident in studies of monozygotic (MZ) and dizygotic (DZ) pairs of twins. In contrast to what would be expected for individuals sharing 100% of their inherited genetic material, the overall pairwise concordance rate of MS in MZ twins have shown to be 25.4% (Willer et al, 2003). Although significantly greater than the DZ concordance rate of 5.4%, and reaching 34% for female MZ after separation by gender, this results still raise questions about what could determine the remaining heritability.

Furthermore, studies have shown decreasing relapse rate and lower disease activity in pregnant women followed by an increase of symptoms during the period immediate after childbirth (Pozzilli & Pugliatti, 2015). One of the most established theories regarding this protective effect of pregnancy on multiple sclerosis is that sex hormones, such as estrogen, released during pregnancy causes immunological changes by directing the development of T helper cells into predominantly the anti-inflammatory Th2 instead of the pro-inflammatory Th1.

Studies like those of monozygotic twins and pregnancy indicates that the fundamental causes of MS are not solely genetic. In this regard, recent research have also focused on environmental and epigenetic factors.

Epigenetic changes include DNA methylation and modifications of histone proteins, of which methylation has been the focus of research due to its stability and that it has been easier to measure in human studies compared to histone modifications (Guintivano & Kaminsky, 2016). It has been shown that DNA methylation could provide a mechanistic explanation regarding the ability for environmental factors to alter genetic expression. This modification occurs in the dinucleotides of cytosine and guanine (CpG), specifically on the 5’ carbon of the cytosine ring.
Genomic regions with a high frequency of CpG sites, called CpG islands, have been of particular scientific interest. The CpG islands have been defined as DNA sequences of at least 200 bp, with a CG content of >50% and a CpG ratio of 0.6 (i.e. 60%). However, these thresholds are quite arbitrary and varying the values have shown to significantly alter analysis results (Illingworth & Bird, 2009). For example, in a study by Takai and Jones (2002) a minimum sequence length of 500 bp, CG content of >55% and CpG ratio of 0.65 was proven to largely exclude false non-CpG island sequences.

Genome-wide studies have been able to show that most epigenetic changes can be regarded as allele specific methylation (ASM) and that those are associated with certain SNPs. Furthermore, studies have revealed consistent ASM patterns in three generations of family members, suggesting an epigenetic inheritance that could in part explain the high disease heritability but few genetic findings observed for several mental disorders (Illingworth & Bird, 2009).

Studies have also shown that environmental factors are able to cause epigenetic changes which then can be transmitted to subsequent generations. Thus, the environment can interact with the genome through epigenetic pattern alterations, influencing the phenotype (Guintivano & Kaminsky, 2016).

In the case of MS, low concordance rates in MZ twins and the fact that the disease mainly are transmitted from mother to offspring and that MS is twice as common in females than males, indicates epigenetic causes (Küçükali et al, 2015).

In a genome-wide DNA methylation study performed by Graves et al. (2014) CD4+ T cells from 30 MS patient and 28 healthy controls were analysed in order to identify methylation changes associated with MS. The study identified a total of 74 CpG sites located in 38 different genes, and especially accumulated within the MHC region. The results also showed high methylation of eight closely located sites in the major MS-associated HLA-DRBI gene, and the study suggests the DRBI*1501 haplotype to be in part responsible for the DNA methylation in this MS risk region. Furthermore, the researchers received distinct methylation signals in regions outside MHC, of which many was located at genes previously associated with MS risk.

To date, a total of 44 environmental risk factors have been identified as associated with multiple sclerosis (Belbasis et al, 2015). However, only two of these, namely the Epstein-Barr virus and smoking, are supported by consistent and reliable results. Although sun exposure and vitamin D deficiency also have been frequently identified as a risk factor for MS, according to Belbasis et al. the evidence for this have been weak.

3. Materials and methods
This section addresses the practical aspects of the study, such as the bioinformatics tools used for data processing. Regarding the methodology that has been drawn up for the upcoming project, a brief description is found in the beginning of this section, which understandable is not covered in the results section later in this report.

3.1 Brief comment on the SNEMS project methods
The planned project (with the name SNEMS) is performed through a collaboration between two research groups at Linköping University, situated at the Department of Physics, Chemistry and
Biology (IFM), led by Mika Gustafsson, and the Department of Clinical and Experimental Medicine (IKE).

The aim of the planned project is to use a multi-omics approach to predict upstream regulators of MS, for the purpose of individualised early treatment using systems oriented tools. For this purpose, the SNEMS project aims to utilize the benefits offered by next-generation sequencing, in order to analyse a selected set of specific SNPs associated with MS through target SNP genotyping with customized sequencing libraries, and to relate these results to previously obtained gene expression data of unstimulated, stimulated and stimulated + treated CD4+ T-cells.

The starting material for this work consists of purified DNA from blood sample CD4+ T cells. These samples originate from 30 MS patients and 14 healthy controls. The plan is to measure the 761 SNPs that meet the threshold of p<10^{-5}, as well as their 332 LD partner SNPs using the TruSeq technology provided by the sequencing company Illumina®. In this way, targeted sequencing on customised microarrays could be conducted, using 250 bp amplicons, which are short pieces of DNA created by polymerase chain reaction (PCR) based on the SNPs of interest. The amplicons make it possible to detect the particular SNPs during genotyping, and they are planned to be designed in Illumina’s web tool DesignStudio.

In addition, there is also of major interest to analyse methylation pattern, and to relate these results to the genotyped SNPs in order to gain further insight into their relationship. In this regard, possible suitable methylation sequencing kits from Illumina® have been identified, although not included in this report.

3.2 Preliminary study methods

3.2.1 Selection of significant SNPs
In GWAS in general, after test for association between the frequency of common variants and a given phenotype, SNPs that exceed a fixed genome-wide threshold for association are selected for further analysis. SNPs with a p value < 5×10^{-8} are usually considered as significant (Gibson, 2010). A list of SNPs of interest was acquired from the supervisor. This list contained a total of 443219 SNPs denoted by an id-number (rs-number), and this data is based on the results from the comprehensive GWAS as conducted by IMSGC (Sawcer et al., 2012). As previously mentioned, a total of 102 SNPs significantly associated with MS were identified in the IMSGC study, using a threshold p-value of p<10^{-8}. For the preliminary study presented in this report, instead SNPs were selected with respect to p<10^{-5}, in order to include more SNPs with weaker level of MS-association. This is because, as described above, multiple sclerosis is considered as a complex disease where many genetic variations each are believed to contribute with a small amount of influence on MS susceptibility. In this regard, p<10^{-5} has proven to provide the greatest enrichment of regulatory SNPs from this list, while a less restrictive threshold value, such as p<10^{-3}, generates too many false positives.

3.2.2 Validation of linkage disequilibrium for the SNPs
The SNPs chosen for the study according to the above description may not constitute the only cause of MS disease. These SNPs have in previous studies been obtained through analysis of microarrays, where determination of LD has not been made. Therefore, in this study, it was
evaluated whether there was loci in the SNPs surroundings that showed LD with them. In this way it was ensured that additional loci of possible importance for MS were identified, and this could provide useful information for the construction of the amplicons in the planned genotyping experiment, as described in the section below.

For this evaluation, the web-based tool SNAP v2.2 from the Broad Institute was used (Johnson et al, 2008). This bioinformatics toolkit relies on data from the International HapMap Project and the 1000 Genomes Project, and it provides the ability to input SNP data for analysis of LD.

To begin with, analysis of pairwise LD among the 761 SNPs was performed, as an initial evaluation of linkage disequilibrium within this selected group of high association SNPs. The linkage disequilibrium threshold for the SNPs to meet was set to $r^2 = 0.8$ and a distance limit was set to 500 kbp.

Second, a search for proxy SNPs based on linkage disequilibrium and physical distance was conducted using the “Proxy Search” in SNAP. Also here, the linkage disequilibrium threshold was set to $r^2 = 0.8$ and the distance limit was 500 kbp.

### 3.2.3 Mapping SNPs to chromosomal positions

Mapping of SNPs to locations on the genome was conducted using the web based software Ensembl’s Biomart from *The International Genome Sample Resource* (IGSR) (Cunningham et al, 2015), based on the human reference genome hg19. In this way chromosomal positions could be obtained for each SNP. Although such data had already been obtained in the previously conducted GWAS and was present in the original SNP list, there was of interest to validate these data by conducting another mapping. The mapping was made for the SNPs as selected based on the $p<10^{-5}$ threshold. In addition, chromosomal mapping was also performed on the new SNPs identified through the LD-analysis described above.

Information on chromosomal positions are essential for the construction of the genotyping analysis, and therefore of importance for the upcoming project.

### 3.2.4 Examination of regulatory properties

In the original list of SNPs a score value describing the SNPs association with regulatory processes had also been specified, using the software *RegulomeDB* (Boyle et al, 2012), based on the human reference genome hg19. This describes which SNPs that can be regarded as expression quantitative trait loci (eQTLs), which are genomic loci that have a role in the level of expression of mRNAs. The software is based on a scoring system with 14 scoring values, each representing current supporting regulatory data. The lower values indicates processes as transcription factor binding, while higher values also are supported by results showing eQTL features.

Using the *RegulomeDB* software, regulatory scores was also obtained for the additional SNPs identified through the LD-analysis as described earlier.

### 3.2.5 Allele frequencies and gene identification

In order to evaluate which variants that could be regarded as common or rare, allele frequencies was analysed. This was conducted using the R programming language with the analysis package *rsnps* (Chamberlain & Ushey, 2015). In this way, data on mean allele frequencies (MAF) based on the global population could be obtained.
Although not included in this study, there should be possible to relate these results to data on odds ratio (the odds to find SNPs in MS patients compared to controls) in order to obtain allele frequencies for MS patients. This would also be convenient in the examination of SNPs rarely found in the human population and therefore could be considered as inadequate to include in a sequencing analysis.

Furthermore, genes in which the SNPs are located was also able to be identified through this analysis package, which was useful to map the new SNPs to genes.

3.2.6 Mapping genes to biological pathways
All genes containing any of the examined SNPs were analysed in regard to known biological pathways, using the software WEB-based Gene SeT AnaLysis Toolkit (WebGestalt) (Wang et al, 2013). The genes with most number of mapped SNPs was also more closely examined through searches in the National Center for Biotechnology Information (NCBI) gene database, the UniProt Knowledgebase (UniProtKB) and the scientific literature.

4. Results
Figure 2 shows a manhattan plot depicting the distribution of p-values among the 22 human autosomes (body chromosomes) for the original GWAS SNP data list. In this plot the y-axis shows the negative base 10 logarithms of the p-values, while the x-axis shows chromosomal positions. The blue line intersecting the chart represents the p-value threshold used to select significant SNPs. As can be seen in the manhattan plot, a large amount of highly significant SNPs are located on chromosome 6, which is in line with current knowledge considering the immunological properties and polymorphic characteristics of the HLA region.

Figure 2: Manhattan plot of p-value distribution among the 22 human autosomes. The y-axis shows the negative logarithms (base 10) of p-values, and the x-axis shows chromosome positions. The plot was constructed using the R-programming package "qqman".
A total of 761 SNPs met the association p-value requirement of $p < 10^{-5}$, providing the starting data material for this preliminary study.

4.1 Examination of linkage disequilibrium

The LD analysis yielded about 1900 links with $r^2 > 0.8$ between SNPs. Among these, a total of 332 new SNPs could be identified as not present on our list of high association SNPs. However, 39 of those was found in the large original SNP list, but of course below the p-value threshold. All these additional SNPs were included in the continued data analysis, giving a total of 1093 SNPs of potential interest for the upcoming project.

The remainder of the results indicates LD between the 761 analysed SNPs themselves. However, the major purpose of this analysis was to identify additional linked SNPs, and therefore further study of the intra-linkages of the 761 SNPs are not included in this study, but may be of interest for future analysis.

4.2 Mapping of SNPs to chromosomal positions

According to the mapped chromosome positions that had been analysed in the previous GWAS, and as expected, a majority of the 761 high associated SNPs that met the p-value threshold was indeed located in the MHC region on chromosome 6 (530 out of 761, nearly 70%) as is particularly evident in Figure 2 above.

However, although chromosomal positions had already been obtained, it was of interest to validate these data by performing another mapping. The results was consistent with the data, but in addition it also yielded multiple positions for many individual SNPs. The reasons is that the MHC region, in which most of these SNPs are located, is characterized by a large sequential variation, and the additional positions identified by the software corresponds to the eight MHC haplotypes previously described. For the continued data analysis, chromosome positions derived from the reference haplotype PGE was used, which was the same as those in the original SNP list. However, it may be appropriate to take the other MHC haplotypes into consideration in the planning of the upcoming project, which is further discussed later in this report.

The additional SNPs identified through the LD analysis showed a more even representation of chromosomes. Chromosome 6 was still found among those chromosomes with most SNPs, however only a few of the new SNPs was located within the MHC region.

Overall, chromosome 6 contains little more than half of the 1093 identified SNPs, as seen in table 1 and figure 3 below.

It could be appropriate to perform a similar selection as above, but with a p-value threshold of $10^{-8}$, in order to investigate how the chromosomal distribution of significant SNPs would differ.
Table 1: Chromosomes and the amount of SNPs located at each of these.

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Number of SNPs</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>97</td>
<td>8.9 %</td>
</tr>
<tr>
<td>2</td>
<td>17</td>
<td>1.6 %</td>
</tr>
<tr>
<td>3</td>
<td>89</td>
<td>8.1 %</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>1.5 %</td>
</tr>
<tr>
<td>5</td>
<td>27</td>
<td>2.5 %</td>
</tr>
<tr>
<td>6</td>
<td>587</td>
<td>53.7 %</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>0.5 %</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>0.4 %</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>0.4 %</td>
</tr>
<tr>
<td>10</td>
<td>22</td>
<td>2.0 %</td>
</tr>
<tr>
<td>11</td>
<td>16</td>
<td>1.5 %</td>
</tr>
<tr>
<td>12</td>
<td>22</td>
<td>2.0 %</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>0.2 %</td>
</tr>
<tr>
<td>14</td>
<td>32</td>
<td>2.9 %</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>0.0 %</td>
</tr>
<tr>
<td>16</td>
<td>66</td>
<td>6.0 %</td>
</tr>
<tr>
<td>17</td>
<td>20</td>
<td>1.8 %</td>
</tr>
<tr>
<td>18</td>
<td>3</td>
<td>0.3 %</td>
</tr>
<tr>
<td>19</td>
<td>29</td>
<td>2.7 %</td>
</tr>
<tr>
<td>20</td>
<td>9</td>
<td>0.8 %</td>
</tr>
<tr>
<td>21</td>
<td>0</td>
<td>0.0 %</td>
</tr>
<tr>
<td>22</td>
<td>26</td>
<td>2.4 %</td>
</tr>
<tr>
<td>Tot:</td>
<td>1093</td>
<td>100 %</td>
</tr>
</tbody>
</table>

4.3 Regulatory scoring of SNPs
Among the 761 high associated SNPs 195 (25.6%) were valuated as eQTLs, while among the 332 LD identified new SNPs 39 (11.7%) were valuated as eQTLs. In total there were 234 eQTLs, of which 134 were located within genes and 100 in non-genic regions.

4.4 Allele frequencies
For the 761 high association SNPs the average MAF was about 0.28, with the highest frequency at 0.4996 for rs12599402 and the lowest at 0.006 for rs17404424. 83 SNPs (10.9%) had a MAF below 0.1 and 23 SNPs (3%) was found below a MAF value of 0.05.

For the 332 new SNPs the average MAF was 0.32, with the highest frequency at 0.4992 for rs12025416 and the lowest at 0.0369 for rs4809339. 13 SNPs (3.9%) had a MAF below 0.1 and just one SNP (0.3%) was found below a MAF value of 0.05.

4.4 Examination of identified SNP-containing genes
Among the MS associated SNPs about half of them (373 out of 761) was found to be located within genes. Among the additional LD identified SNPs a majority (260 out of 332) were located in genes. Together, this gave 633 genic SNPs, corresponding to 196 unique genes. Table 2 shows all the genes and their constituent number of SNPs. The top nine genes, with 10 or more SNPs, are highlighted green in the table. The CLEC16A gene had the highest amount of mapped SNPs, followed by TAP2, MANBA, BACH2, MUC22, LIC00271, C6orf10, AHI1 and CCHCR1.
Table 2: All the unique genes where analysed SNPs have been mapped. The first nine genes, which contains 10 or more SNPs, are highlighted in green.

<table>
<thead>
<tr>
<th>Gene</th>
<th>SNPs</th>
<th>Gene</th>
<th>SNPs</th>
<th>Gene</th>
<th>SNPs</th>
<th>Gene</th>
<th>SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLEC16A</td>
<td>35</td>
<td>LOC105378922</td>
<td>4</td>
<td>MICA</td>
<td>2</td>
<td>BTN3A3</td>
<td>1</td>
</tr>
<tr>
<td>TAF2</td>
<td>16</td>
<td>TNIP2</td>
<td>4</td>
<td>BATF</td>
<td>2</td>
<td>MALI</td>
<td>1</td>
</tr>
<tr>
<td>MANBA</td>
<td>15</td>
<td>EXTL2</td>
<td>4</td>
<td>RNFLP1</td>
<td>2</td>
<td>PLCL2</td>
<td>1</td>
</tr>
<tr>
<td>BACH2</td>
<td>14</td>
<td>VARS2</td>
<td>4</td>
<td>CANA1S</td>
<td>2</td>
<td>HSD17B8</td>
<td>1</td>
</tr>
<tr>
<td>MUC22</td>
<td>14</td>
<td>BAG6</td>
<td>4</td>
<td>CD86</td>
<td>2</td>
<td>AGPAT1</td>
<td>1</td>
</tr>
<tr>
<td>LINC00271</td>
<td>14</td>
<td>BTNL2</td>
<td>4</td>
<td>PDE4C</td>
<td>2</td>
<td>IL12A</td>
<td>1</td>
</tr>
<tr>
<td>C6orf10</td>
<td>12</td>
<td>LINC01146</td>
<td>4</td>
<td>SLC30A7</td>
<td>2</td>
<td>POGUT1</td>
<td>1</td>
</tr>
<tr>
<td>AHI1</td>
<td>10</td>
<td>TAP1</td>
<td>3</td>
<td>HLA-DMB</td>
<td>2</td>
<td>TUBD1</td>
<td>1</td>
</tr>
<tr>
<td>CHIC1R1</td>
<td>10</td>
<td>PSM8</td>
<td>3</td>
<td>LOC283587</td>
<td>2</td>
<td>LOC10056047</td>
<td>1</td>
</tr>
<tr>
<td>CSF2R8</td>
<td>9</td>
<td>TRIM31</td>
<td>3</td>
<td>ZNF767</td>
<td>2</td>
<td>IL7R</td>
<td>1</td>
</tr>
<tr>
<td>NDFIP1</td>
<td>9</td>
<td>EAF2</td>
<td>3</td>
<td>CDC37</td>
<td>2</td>
<td>HCP5</td>
<td>1</td>
</tr>
<tr>
<td>EVI5</td>
<td>9</td>
<td>RP54K8B1</td>
<td>3</td>
<td>ABH16A</td>
<td>2</td>
<td>IP6K3</td>
<td>1</td>
</tr>
<tr>
<td>NOTCH4</td>
<td>9</td>
<td>KIF12B</td>
<td>3</td>
<td>TCF19</td>
<td>2</td>
<td>PPR1R10</td>
<td>1</td>
</tr>
<tr>
<td>PSORS1C1</td>
<td>8</td>
<td>HLA-DQA2</td>
<td>3</td>
<td>PTT2</td>
<td>2</td>
<td>SLC15A2</td>
<td>1</td>
</tr>
<tr>
<td>LLRD34</td>
<td>8</td>
<td>LOC105371082</td>
<td>3</td>
<td>CDK4</td>
<td>2</td>
<td>CXCR5</td>
<td>1</td>
</tr>
<tr>
<td>STAT3</td>
<td>8</td>
<td>TCC3</td>
<td>3</td>
<td>PRRC2A</td>
<td>2</td>
<td>SLC6A16</td>
<td>1</td>
</tr>
<tr>
<td>FAM69A</td>
<td>8</td>
<td>BRD2</td>
<td>3</td>
<td>DDR1</td>
<td>2</td>
<td>PRM2</td>
<td>1</td>
</tr>
<tr>
<td>PLEK</td>
<td>7</td>
<td>HLA-DPA1</td>
<td>3</td>
<td>ELMO1</td>
<td>2</td>
<td>SP110</td>
<td>1</td>
</tr>
<tr>
<td>PXT1</td>
<td>7</td>
<td>MMEL1</td>
<td>3</td>
<td>MAST3</td>
<td>2</td>
<td>LOC105369325</td>
<td>1</td>
</tr>
<tr>
<td>GOLGB1</td>
<td>7</td>
<td>HLA-DPB1</td>
<td>3</td>
<td>PSMB9</td>
<td>2</td>
<td>LST1</td>
<td>1</td>
</tr>
<tr>
<td>MAPK1</td>
<td>7</td>
<td>MOG</td>
<td>3</td>
<td>METTL1</td>
<td>2</td>
<td>LOC105369627</td>
<td>1</td>
</tr>
<tr>
<td>IL2RA</td>
<td>7</td>
<td>HLA-DPB2</td>
<td>3</td>
<td>AGAP2</td>
<td>2</td>
<td>LG658</td>
<td>1</td>
</tr>
<tr>
<td>Iqcb1</td>
<td>6</td>
<td>ARHGAAP1</td>
<td>3</td>
<td>DKKL1</td>
<td>2</td>
<td>ARHGEF25</td>
<td>1</td>
</tr>
<tr>
<td>FCRL3</td>
<td>6</td>
<td>TNFRSF1A</td>
<td>3</td>
<td>PVR</td>
<td>2</td>
<td>TBC1D3P1-DHX40P1</td>
<td>1</td>
</tr>
<tr>
<td>ETV7</td>
<td>6</td>
<td>HCG9</td>
<td>3</td>
<td>AGER</td>
<td>2</td>
<td>C80D</td>
<td>1</td>
</tr>
<tr>
<td>HLA-F-AS1</td>
<td>6</td>
<td>TRIM40</td>
<td>3</td>
<td>HCLS1</td>
<td>2</td>
<td>TIMMDC1</td>
<td>1</td>
</tr>
<tr>
<td>IL12A-AS1</td>
<td>6</td>
<td>POUSF1</td>
<td>3</td>
<td>LOC100294145</td>
<td>2</td>
<td>PSORS1C3</td>
<td>1</td>
</tr>
<tr>
<td>MR0H3P</td>
<td>6</td>
<td>PPM1F</td>
<td>3</td>
<td>LOC105370548</td>
<td>2</td>
<td>HLA-DRA</td>
<td>1</td>
</tr>
<tr>
<td>VMP1</td>
<td>5</td>
<td>DPH5</td>
<td>3</td>
<td>GPX6</td>
<td>1</td>
<td>DDX39B</td>
<td>1</td>
</tr>
<tr>
<td>MYNN</td>
<td>5</td>
<td>TAGAP</td>
<td>2</td>
<td>LYG66</td>
<td>1</td>
<td>TNFSF14</td>
<td>1</td>
</tr>
<tr>
<td>CD6</td>
<td>5</td>
<td>LOC105378081</td>
<td>2</td>
<td>SOCS1</td>
<td>1</td>
<td>PVT1</td>
<td>1</td>
</tr>
<tr>
<td>HLA-DOA</td>
<td>5</td>
<td>IFI30</td>
<td>2</td>
<td>B4GALNT1</td>
<td>1</td>
<td>TNX8</td>
<td>1</td>
</tr>
<tr>
<td>OS9</td>
<td>5</td>
<td>ZBTC46</td>
<td>2</td>
<td>TOP3B</td>
<td>1</td>
<td>C2</td>
<td>1</td>
</tr>
<tr>
<td>LOC100965853</td>
<td>5</td>
<td>SAE1</td>
<td>2</td>
<td>HCG23</td>
<td>1</td>
<td>GPR65</td>
<td>1</td>
</tr>
<tr>
<td>GALT</td>
<td>5</td>
<td>CD58</td>
<td>2</td>
<td>SLC12A5</td>
<td>1</td>
<td>LOC105378083</td>
<td>1</td>
</tr>
<tr>
<td>COL11A2</td>
<td>5</td>
<td>HLA-DQA1</td>
<td>2</td>
<td>NCOA5</td>
<td>1</td>
<td>ALPK2</td>
<td>1</td>
</tr>
<tr>
<td>HCG17</td>
<td>5</td>
<td>MPV1L2</td>
<td>2</td>
<td>LTA</td>
<td>1</td>
<td>RING1</td>
<td>1</td>
</tr>
<tr>
<td>HLA-DQB</td>
<td>5</td>
<td>UBSH3B</td>
<td>2</td>
<td>KEAP1</td>
<td>1</td>
<td>MICB</td>
<td>1</td>
</tr>
<tr>
<td>TRIM10</td>
<td>4</td>
<td>GPX5</td>
<td>2</td>
<td>DLEU1</td>
<td>1</td>
<td>GF11</td>
<td>1</td>
</tr>
<tr>
<td>CYP24A1</td>
<td>4</td>
<td>RPS18</td>
<td>2</td>
<td>KIAA1683</td>
<td>1</td>
<td>IL22RA2</td>
<td>1</td>
</tr>
<tr>
<td>RGS14</td>
<td>4</td>
<td>LOC100507463</td>
<td>2</td>
<td>TSI FM</td>
<td>1</td>
<td>RPPA2</td>
<td>1</td>
</tr>
<tr>
<td>DPCR1</td>
<td>4</td>
<td>SP140</td>
<td>2</td>
<td>HCG26</td>
<td>1</td>
<td>WDR46</td>
<td>1</td>
</tr>
<tr>
<td>TNFRSF14</td>
<td>4</td>
<td>HCG22</td>
<td>2</td>
<td>ZFPM1L1</td>
<td>1</td>
<td>LOC115110</td>
<td>1</td>
</tr>
<tr>
<td>ITPR3</td>
<td>4</td>
<td>LYY66C6</td>
<td>2</td>
<td>OR14J1</td>
<td>1</td>
<td>C6orf47</td>
<td>1</td>
</tr>
<tr>
<td>ZMIZ1</td>
<td>4</td>
<td>NCF4</td>
<td>2</td>
<td>LOC400499</td>
<td>1</td>
<td>GGNBP1</td>
<td>1</td>
</tr>
<tr>
<td>ILDR1</td>
<td>4</td>
<td>ACTRRT3</td>
<td>2</td>
<td>EGFL8</td>
<td>1</td>
<td>MUC21</td>
<td>1</td>
</tr>
<tr>
<td>Clilorf106</td>
<td>4</td>
<td>HCG27</td>
<td>2</td>
<td>GNL1</td>
<td>1</td>
<td>LOC285626</td>
<td>1</td>
</tr>
<tr>
<td>ARHGEF3</td>
<td>4</td>
<td>MERTK</td>
<td>2</td>
<td>PDE4A</td>
<td>1</td>
<td>SKIV2L</td>
<td>1</td>
</tr>
<tr>
<td>TMEM39A</td>
<td>4</td>
<td>NFKB1L1</td>
<td>2</td>
<td>HLA-DQB1</td>
<td>1</td>
<td>LOC105370612</td>
<td>1</td>
</tr>
</tbody>
</table>

When looking only at eQTL-SNPs the 134 genic ones mapped to 76 unique genes, while the remaining 100 non-genic eQTL-SNPs were located close (<60kb) to 46 unique genes.
4.5 Mapping genes to biological pathways

Of the 196 genes inserted into Webgestalt, 185 could be mapped to 185 unique Entrez Gene IDs, while 11 were mapped to multiple Entrez Gene IDs or could not be mapped to any at all. Thus, the software analysis was based upon the 185 unique Entrez Gene IDs.

Figure 4 summarizes the different biological processes obtained for the analysed gene set. Each category is represented by a bar whose height corresponds to the number of genes associated with that particular biological process. According to the result, biological regulation constitute the dominant process (91 of 185 genes), followed closely by metabolic processes, response to stimulus, multicellular organismal processes, cell communication and developmental processes. Figure 5 further clarifies the main areas of biological processes in which the analysed genes are active. Here it is evident that the dominant areas (highlighted in red) consists of immune system processes and their regulation, immune response and regulation of these, and antigen processing and presentation.

Figure 4: A bar chart of different biological processes which the genes have been mapped to. The height of each category corresponds to the number of genes associated with that particular process. A gene may therefore be associated with several biological processes.
Figure 5: Directed Acyclic Graph (DAG) containing gene ontology (GO) categories of biological processes. GO categories labelled red represents enriched categories and therefore biological areas of possible importance.

Enrichment analysis with Webgestalt was also performed on only genes related to the eQTL-SNPs, genic and non-genic respectively. In both cases biological regulation was indicated as
the major biological process in both cases, similar to the results in figure 4, as well as response to stimulus and developmental processes.

Through disease association analysis, the genes related to genic eQTL-SNPs was identified as particularly associated with autoimmune diseases (HLA-genes, CLEC16A, TAP2, C6orf10), disease susceptibility and immune system diseases. Genes related to non-genic eQTL-SNPs were mainly associated with immune system diseases (most of them HLA-genes), autoimmune diseases, and also virus diseases and infection.

5. Discussion, conclusions and ideas for the planned project

As previously stated, this study is to be considered as a start-up study conducted in order to investigate the scientific basis, and to examine the data set, prior to a larger upcoming research project. Therefore, the most important results is still to be expected in subsequent research work, such as the analysis of methylation patterns. However, the initial study presented in this paper have provided a first informative view of the material to be analysed in the research project, as well as in some extent the prerequisites for its implementation.

Examination of allele frequencies was conducted in order to evaluate whether there were SNPs with notably low values in the data set, which thus could be unsuitable for the genotyping analysis later in the SNEMS project, where a total of 44 persons are included in the study.

The majority of SNPs exhibited relatively high mean allele frequencies (MAFs) in the global population, with a mean of 0.28 for the 761 highly MS associated SNPs and 0.32 for their 332 LD partner SNPs, i.e. these single nucleotide variations are estimated to be found in 28% and 32% of the human population respectively.

Allele frequencies were thus generally higher for the new LD-identified SNPs, which may be considered somewhat surprising, especially when these are not present in the original data list obtained from the large GWAS. However, this could be regarded as an expected result, because high frequency in healthy persons in relation to MS association gives p-values of less significance.

In the case of the 761 highly MS associated SNPs 83 (10.9%) had a MAF below 0.1 and 23 (3%) was found below a MAF value of 0.05, while for the 332 new SNPs 13 (3.9%) had a MAF below 0.1 and just one SNP (0.3%) was found below a MAF value of 0.05.

SNPs that shows very low allele frequencies, such as rs17404424 (MAF = 0.006), may be important to take in consideration in the design of the genotyping experiment. Due to the relatively low number of patients and controls to be used, there may be a large probability that such rare variants are not present at all and therefore could be regarded as unnecessary to include in the experiment. However, as demonstrated by current scientific literature, rare variants are believed to constitute an important contributing factor for MS susceptibility, giving good reason for the choice to indeed include the low-frequency SNPs identified in the study. In the end, it could be considered as a purely economic issue.

The examination of gene-localized SNPs revealed a number of interesting genes.

The highest number of genic SNPs was found in the CLEC16A gene, 35 in total. The CLEC16A gene is located at chromosome 16p13, encoding the C-type lectin proteins which
functions as regulators of antigen-presenting cells and thereby plays a key role in the immune regulation (van Luijn et al, 2015). This locus has been identified as important for the susceptibility of several autoimmune diseases, such as multiple sclerosis and type I diabetes, although its mechanisms in these human diseases are yet not fully understood.

16 SNPs was located in TAP2, a gene encoding the antigen peptide transporter 2 protein. This protein is located in the endoplasmatic reticulum membrane, and is involved in the transport of antigens as well as the final folding stage of MHC class I molecules (UniProtKB). In a study by Moins-Teisserenc et al. (1995), the results suggested that polymorphisms in the TAP2 gene was associated with MS susceptibility.

MANBA, a protein coding gene on chromosome 4p24, contained 15 of the analysed SNPs. According to the NCBI gene database, MANBA encodes the beta-mannosidase protein, which is present in the lysosomes, and mutations in the gene are associated with neurological diseases. In a follow-up study conducted by the IMSGC (2013a) significant MS association of variations within or near MANBA was revealed, including the rs228614 SNP also present in our data set.

14 SNPs were found in the BACH2 gene, which is located at chromosome 6p15 and encodes transcription regulator protein BACH2 (NCBI gene database). The BACH2 gene is involved as an anti-inflammatory factor in the immune system regulation, required for adequate formation of T_{REG} cells and acts as an important regulator of CD4^{+} T-cell differentiation (Roychoudhuri et al, 2013). In a study by Perga et al. (2015) it could be confirmed that BACH2 transcripts are reduced in blood cells from MS patients, which reinforces the view of MS as a disease not only resulted from an over-reactive immune system but also from the reduction of regulatory functions.

MUC22, a cell membrane associated gene encoding the protein mucin-22, contained 14 of the analysed SNPs. In a study by Hijikata et al. (2011) the MUC22 was mapped to a mucin gene cluster at chromosome 6p21.3, and also revealed expression of the gene in placenta, testis and lungs. However, there seems to be little information regarding the biological function of the MUC22 gene.

LINC00271 also contained 14 of the analysed SNPs. This is a long non-coding gene located on chromosome 6p23.3. No further information has been found regarding this gene. However, many long non-coding RNAs possess regulatory properties and recent research have suggested that they may play a major role in fundamental biological processes as well as in the etiology of psychiatric disorders (Lingjun et al, 2016).

12 SNPs were mapped to the C6orf10 gene, which is located on chromosome 6p21.3 in the MHC region and encodes the uncharacterized protein C6orf10 (NCBI gene database).

10 SNPs were mapped to AHI1, a gene located on chromosome 6p23.3 and encoding the protein jouberin (NCBI gene database). In a study by Lin et al. (2015a) six eQTL SNPs located in five eQTL genes were significantly associated with MS. Interestingly, of those SNPs only two were found in our 1093 SNP list, namely rs11154801 in the AHI1 gene and rs1062158 in the gene NDFIP1. New SNPs previously not identified in our list was rs3095329, rs9469220, rs2647046 and rs7194. These SNPs have not been analysed in this study, but it may be appropriate to include them in the forthcoming project.

Another 10 SNPs were mapped to CCHCR1, a gene on chromosome 6p21.3 encoding coiled-coil alpha-helical rod protein 1 (NCBI gene database), which can be related to cell differentiation and protein transport (UniProtKB). The CCHCR1 gene has also been associated with high expression in MS-related cells (Lin et al, 2015b).
In conclusion, most of these nine genes have received a significant association with MS in the scientific literature. It may seem odd that the strongest MS associated gene HLA-DRB1, as described in the theoretical background section, is not present among the identified genes. However, this could be due to the type of microarray used in the GWAS from which the initial data for this study have been obtained, and therefore not all possible SNPs have been covered. Another explanation could be that the HLA-DRB1 gene is affected by mainly non-genic regulatory SNPs. Indeed, among the non-genic eQTL-SNPs described in the results section, one such SNP had HLA-DRB1 as its nearest gene (about 20kb away).

The fact that 100 regulatory MS-associated SNPs are located outside genes is consistent with current scientific knowledge, where the human genome is known to contain a lot of regulatory DNA in addition to protein coding regions (Woolfe et al, 2004).

It would also be convenient to examine whether the genic SNPs are located in the coding parts (exons) of the genes. There has not been enough time to do this within the timeframe of this study, and therefore such analysis may be performed in future studies.

Initially, the aim of this study also included a first minor examination of methylation data. This was intended to act as a comparative example of the relationship between genetic and epigenetic factors in multiple sclerosis. Unfortunately, the study’s time frame required some adjustments to be made, and the methylation data examination was one of the things that had to be excluded. However, study of methylation patterns to be compared with the genotyping results is still planned to be conducted as part of the SNEMS project.

It seems clear that the MHC (HLA) region, besides its crucial role in human diseases, constitute a highly complex part of the human genome, and makes sequencing this region a difficult task.

As previously described the mapping of the high MS association SNPs to chromosomal positions yielded multiple haplotype-related positions for many individual SNPs, as a result of the polymorphic characteristics of the MHC region. At the moment, amplicons for the sequencing SNEMS project will be based on chromosomal positions derived from the reference haplotype PGE.

However, it would be of interest to be able to also include the other SNP-haplotypes. Notably, it has been shown that the use of SNP-haplotypes explains more of MS heritability compared to if just single-SNP analysis is used. SNP-haplotypes are thus useful for fine-mapping genetic regions identified by previous GWAS, and might even lead to the identification of new disease-associated loci (Khankhanian et al, 2015).

In order to make the amplicons able to detect the same SNP in all the haplotypes, the sequence around the SNP need to be conserved between the haplotypes. The question to address is whether amplicons of 250 bps is enough to cover all the eight variants of the MHC region. Therefore, there is a need to further examine the genetic environment around the SNPs, preferably 125 bps left and right of the SNP giving a total span of 250 bps. It is essential that the region around a given SNP is more or less identical among the different haplotypes in order for the genotyping experiment to be implemented as planned.

At the time of writing, new data have been received from another comprehensive study by IMSGC (2013). This data includes approximately 5500 SNPs of significance for MS (according to p-value <10^-5). A brief examining comparison between the IMSGC data set and the results
presented in this report showed that our MS-significant SNPs was found here, as well as about one third of their identified LD partners, thus confirmed the reliability of our results.

However, trying to include all these SNPs in the upcoming sequencing experiment entails some issues. Because the maximum number of possible amplicons that can be constructed for the Illumina® sequencing kits are 1500, it is crucial to know whether 250bp sized amplicons are able to cover an appropriate number of all these SNPs. In the brief comparison mentioned above it was evident, unfortunately, that it would need more than 5000 amplicons to achieve this degree of coverage.

In order to address this issue there are at least two solutions that could be useful. Either by lowering the p-value threshold, thereby restricting the number of significant SNPs, or by only focus on eQTL-SNPs, which after all are considered as possible main actors in the onset of autoimmune disease.
6. References

6.1 Literature

Al-Tubaikh J.A. (2010), *Internal Medicine: An Illustrated Radiological Guide*, Springer-Verlag Berlin Heidelberg. [Chapter 2.8, Multiple sclerosis and other demyelinating diseases, pp. 84-90]


Hofker M.H., Fu J. and Wijmenga C. (2014), *The genome revolution and its role in understanding complex diseases*


Horton et al. (2008), *Variation analysis and gene annotation of eight MHC haplotypes: The MHC Haplotype Project*, Immunogenetics, Vol. 60, pp. 1-18


IMSGC (2013a), *MANBA, CXCR5, SOX8, RPS6KB1 and ZBTB46 are genetic risk loci for multiple sclerosis*, BRAIN, Vol. 136, pp. 1778-1782


Lill C.M. (2014), *Recent advances and future challenges in the genetics of multiple sclerosis*, Front Neurol., Vol. 5, Article 130


Sawcer et al. (2014), Multiple sclerosis genetics, Lancet Neurol., Vol. 13, pp. 700-709


Slatkin M. (2008), Linkage disequilibrium – understanding the evolutionary past and mapping the medical future, Nature Reviews, Vol. 9, pp. 477-485

Sweden Medical Products Agency (Läkemedelsverket), Läkemedelsbehandling av multipel skleros (MS), [website], accessed 2016-04-01


6.2 R programming packages


6.3 Databases

Available: http://www.uniprot.org/help/uniprotkb

National Center of Biotechnology Information (NCBI)
Publikationsstitel / Title
The Complex Genetics of Multiple Sclerosis - A preliminary study of MS-associated SNPs prior to a larger genotyping project

Författare / Author
Simon Söderholm

Sammanfattning / Abstract
Biomedical research have been revolutionized by recent technological advances, both in the fields of molecular biology and computer science, turning the molecular and genetic research into “big data science”. One of the main objectives have been to improve our understanding of complex human diseases. Among those diseases, multiple sclerosis (MS) is considered as one of the most common. MS is a chronic autoimmune disease that cause inflammation and damage to the central nervous system. In this study, a set of bioinformatics analyses have been conducted on SNP data, as an initial step to gain more information prior to an upcoming genotyping project. The results showed extensive regulatory properties for the 761 selected SNPs, which is consistent with current scientific knowledge, and also identified another 332 SNPs in linkage to these. However, during the study some issues have also been identified, which need to be addressed going forward.

Nyckelord / Keywords
Multiple Sclerosis, SNPs, Linkage Disequilibrium, T-cells, MHC, HLA