

From the Department of CLINICAL NEUROSCIENCE
Karolinska Institutet, Stockholm, Sweden

HLA GENETICS IN MULTIPLE SCLEROSIS

Jenny Link



**Karolinska
Institutet**

Stockholm 2012

All previously published papers were reproduced with permission from the publisher.

Published by Karolinska Institutet.

© Jenny Link, 2012

ISBN 978-91-7457-674-0

“Människokroppen är som en buss som kör generna en bit på deras resa i evigheten”
- Johanna Björling

ABSTRACT

Multiple sclerosis (MS) is a chronic disease in which both genetic and environmental risk factors contribute to disease susceptibility. MS patients suffer from inflammatory lesions in the central nervous system which results in demyelination of nerve cells, reduced neuronal activity and finally neurodegeneration. The immune system has a central role in MS pathogenesis and human leukocyte antigen (HLA) molecules are key players. The genes encoding HLA class I and II molecules are highly polymorphic isotypically and allotypically which makes it problematic to identify which variants affect disease susceptibility.

The strongest genetic risk factor for MS is a haplotype of HLA class II alleles, *DRB1*15:01,DRB5*01:01,DQA1*01:02,DQB1*06:02* (below referred to as *DRB1*15*) which increases the risk of MS 3-fold compared with the general Swedish population where the lifetime risk of MS is 0.2%. Our group pioneered the identification of a protective effect of the HLA class I region, by discovering that *HLA-A*02* decreases the risk of MS by 40%. The main focus in this thesis has been to identify additional HLA factors, if any, that influence MS susceptibility.

In **papers I and II** we genotyped 1,784 Swedish and Norwegian MS patients and 1,660 controls, for *HLA-DRB1*, *HLA-A*, *HLA-C*, and eventually also *HLA-B*, and applied several statistical methods, mainly logistic regression analyses. We conclude that, in addition to the roles played by *DRB1*15* and *HLA-A*02*, additional influence on susceptibility is exerted by *HLA-DRB1*01*, *HLA-DRB1*07* and *HLA-B*12*, which are negatively associated with MS and *HLA-B*14* which increases the risk of MS. Analysis based on haplotypes, rather than on alleles, showed that a haplotype carrying *HLA-A*02*, *HLA-C*05* and *HLA-B*12* is markedly protective, reducing the risk of MS 2.4-fold, also outweighing the risk of *HLA-DRB1*15* when present on the same haplotype.

Paper III focuses on a possible interaction between genetic background (*DRB1*15*) and an environmental influence, a month-of-birth effect on MS risk. We demonstrate that patients born in April have a higher risk of being *DRB1*15* positive. On the contrary, patients born in November have a lower risk of being *DRB1*15* positive. We hypothesize that pregnancies exposed to a lower degree of sunlight thus lower levels of Vitamin D, confer an increased risk for a *DRB1*15* positive child to later develop MS.

In **paper IV** the influence of HLA genes on the risk of developing neutralizing antibodies (NAb) to interferon beta (IFN- β) treatment was studied. We show that the risk allele for MS, *HLA-DRB1*15*, is also a risk factor for development of NAb in patients treated with high dose subcutaneously administered IFN- β 1-a, but not for IFN- β 1-b. *DRB1*15* is also a risk factor for developing antibody titers high enough to abolish the effect of treatment. Thus, the genetic risk of NAb varies with IFN- β formulation.

This thesis adds several pieces of information to the large MS genetics puzzle and suggests several roles of HLA genes and molecules that should be further investigated.

LIST OF PUBLICATIONS

- I. **Two HLA class I genes independently associated with multiple sclerosis.**
LINK J, Lorentzen ÅR, Kockum I, Duvefelt K, Lie BA, Celius EG, Harbo HF, Hillert J, Brynedal B.
J Neuroimmunol. 2010 Sep 14;226(1-2):172-6.
- II. **Importance of Human Leukocyte Antigen (HLA) class I and II Alleles on the Risk of Multiple Sclerosis.**
LINK J, Kockum I, Lorentzen ÅR, Lie BA, Celius EG, Westerlind H, Schaffer M, Alfredsson L, Olsson T, Brynedal B, Harbo HF, Hillert J.
Accepted for publication PLoS ONE April 2012.
- III. **HLA-DRB1 and month of birth in multiple sclerosis.**
Ramagopalan SV*, LINK J*, Byrnes JK, Dymment DA, Giovannoni G, Hintzen RQ, Sundqvist E, Kockum I, Smestad C, Lie BA, Harbo HF, Padyukov L, Alfredsson L, Olsson T, Sadovnick AD, Hillert J, Ebers GC.
Neurology. 2009 Dec 15;73(24):2107-11.
* Authors contributed equally
- IV. **Human leukocyte antigen genes and interferon beta preparation influence on risk of developing neutralizing antibodies in multiple sclerosis.**
LINK J, Lundkvist M, Fink K, Hermanrud C, Brynedal B, Fogdell-Hahn A, Kockum I, Hillert J.
Manuscript.

CONTENTS

1	Genetic variations of importance in this thesis.....	1
1.1	Introduction to genetic variation	1
1.1.1	Definitions of concepts in genetics	3
1.2	Patient and control cohorts	5
1.2.1	Study design	5
1.3	Methods of analyzing the data	6
1.3.1	Odds Ratio	6
1.3.2	Null hypothesis, power and significance of a test statistic....	6
2	Human leukocyte antigen.....	8
2.1	HLA molecules.....	8
2.1.1	Class I molecules.....	9
2.1.2	Class II molecules	10
2.1.3	Cross presentation	10
2.2	Genetic variability within HLA.....	10
3	HLA in complex diseases.....	13
3.1	HLA in autoimmune diseases	13
3.1.1	Type 1 Diabetes.....	13
3.1.2	Rheumatoid arthritis.....	14
3.1.3	Systemic lupus erythematosus	14
3.1.4	Celiac disease	14
3.2	Multiple sclerosis.....	14
3.2.1	HLA in multiple sclerosis	15
4	My studies.....	17
4.1	Paper I and II.....	17
4.1.1	The concept of LD within my data	17
4.1.2	Scooped?.....	17
4.1.3	Several reports but only one main point	18
4.1.4	Recursive partitioning didn't work	18
4.1.5	Mimicking the IMSGC paper	19
4.1.6	My way of thinking about logistic regression in paper I and II	19
4.1.7	Haplotypes of allele groups.....	25
4.1.8	Perspectives from papers I and II.....	27
4.2	Paper III.....	27
4.3	Paper IV	28
4.3.1	Patient ascertainment biases.....	30
4.3.2	Genetic impact on NAb development	31
5	Future perspectives.....	33
6	Acknowledgements	35
7	References.....	38

LIST OF ABBREVIATIONS

Ab	Antibody
APC	Antigen presenting cell
Bab	Binding antibody
Bp	Base pair
CNS	Central nervous system
CSF	Cerebrospinal fluid
DNA	Deoxyribonucleic acid
ELISA	Enzyme-linked immunosorbent assay
FDR	False discovery rate
GWAS	Genome wide association study
HLA	Human leukocyte antigen
MS	Multiple sclerosis
IFN- β	Interferon beta
IMSGC	International multiple sclerosis genetics consortium
KIR	Killer cell immunoglobulin-like receptor
LD	Linkage disequilibrium
Mb	Megabase
MBP	Myelin basic protein
MOG	Myelin oligodendrocyte glycoprotein
MRI	Magnetic resonance imaging
MxA	Myxovirus resistance protein A
NAb	Neutralizing antibody
OR	Odds ratio
PCR	Polymerase chain reaction
PLC	Peptide loading complex
PLP	Proteolipid protein
PPMS	Primary progressive MS
RAM	Random access memory
RNA	Ribonucleic acid
RRMS	Relapsing remitting MS
SNP	Single nucleotide polymorphism
SPMS	Secondary progressive MS
T1D	Type 1 diabetes
T1DGC	Type 1 diabetes genetics consortium
TRU	Tenfold reduction units

1 GENETIC VARIATIONS OF IMPORTANCE IN THIS THESIS

To study genetic variations associated with complex disease, it is important to understand its definitions and origins. Here, I will give a brief introduction to how variations in the human genome arise, how it spreads in a population and why it is important. This is only a short overview and there are exceptions to the examples presented as well as more complicated underlying mechanisms that I have chosen to not discuss here.

1.1 INTRODUCTION TO GENETIC VARIATION

When the embryo is formed it has two copies of the human DNA, one from the maternal side and one from the paternal side. Each copy consists of 23 chromosomes and the two homologous setups are basically the same (I here disregard the difference between the X and Y chromosomes, since it does not matter in this respect) although there are small differences between them. One base here and there is different while all the surrounding bases are the same between two homologous chromosomes. The human genome normally has five corner stone bases, adenine (A), cytosine (C), guanine (G), thymidine (T) and urasil (U) in which T is specific to DNA and U to RNA.

Mutation is the term used when a sudden change is introduced in the genome. Mutations can arise from several sources of which errors in DNA replication is one that is self-inflicted, irradiation and viruses are two extraneous sources out of many. DNA replication occurs when a cell prepares for division, e.g. in the developing embryo or when new skin cells are formed. As the cells start to divide in the new individual, more and more mutations arise in the DNA. If the mutations arise in cells that later produce germline cells they will have a chance to be carried on to the next generation. There are cellular mechanisms that identify and correct mutations in the replication process, and cells that do not function correctly are normally eliminated.

Some parts in the genome are more prone to vary than others. Regions where mutations frequently arise are called “mutation hotspots”, and so called “conserved regions” tend to have lower mutation rates. Conserved regions are very similar between species and does often contain some kind of evolutionary important segment that was not beneficial for the individual if altered. This process of non-selection is sometimes referred to as “purifying selection”, and this process keeps new mutations in this region on a low level in the species.

It is important to distinguish the difference between an individual’s DNA and the DNA in the population i.e. all individuals taken together. One specific individual has only two copies of DNA which is then altered within each cell during its lifetime. This results in a mosaicism within the body although the frequency of each mutation is very low, most probably only present in one or very few cells. All these mutations disappear with us when we die unless some were present in a germline cell resulting in a child.

Then the mutations will be present in all cells of the child. This child will have a chromosome from one parent that have one set of mutations and one from the other where other mutations are present. Therefore, when this individual in itself has a child, there will be a 50% chance that the paternal mutations are carried on to the next generation, and 50% for the maternal mutations at a specific location in the genome. By these means, the chance of spreading the mutations through the generations alters the proportion of the population that has the mutation over time. This random event of a mutation being carried on to the next generation is called genetic drift. DNA carried on to the next generation is a sample of the former generation and can thereby alter the population frequency of every base pair in the DNA.

If a mutation of a specific base becomes common in the population it is called a single nucleotide polymorphism (SNP). A SNP is defined as a variation in the population at a specific base pair position with a minor allele frequency of more than 1%. SNPs are generated over time by mutation of the ancestral allele, which is the original allele in the former generations. The SNP is said to have two or more alleles (variants, e.g. A and G) of which one is the ancestral allele and the others are the new alleles. Normally only one alteration of the ancestral allele is present in the population but there are SNPs with three or even four alleles.

There are several factors that may increase the frequency of an allele in the population. Migration of individuals is one such event. Migration is seldom random, usually whole families or closely related individuals move together, thereby increasing the allele frequency of the variants they harbor into the new population. Another event is when a new pathogen strikes a population. If a specific variant of DNA is less beneficial for survival, this allele will decrease in frequency in the population. If there is a selective advantage for a specific allele, this will increase in frequency within the population. These two types of narrowing the allelic diversity of a population are often referred to as a “genetic bottleneck”.

Genes are built up mainly by introns and exons and several other elements, such as promoters. Introns and exons are both transcribed but only exons form the spliced mRNA that translates into a protein. Promoters are the starting sequences for the transcription and are situated in front of the genes. The transcription complex assembles at the promoters upon transcription. The RNA that is formed is then processed in various ways and forms the mRNA which is either functional in itself or is used as a blueprint of a protein. The ribosome assembles the protein by joining amino acids according to the mRNA blueprint. The process of transcription of DNA into RNA and translation of RNA into protein is tightly regulated on all levels. There are signals within the cell that induce or silence transcription, ways of regulating the amount of RNA, and pathways to degrade mRNAs and proteins fast or even protect the protein from degradation. We all carry multiple new mutations, but only a small fraction cause any detectable phenotype.

Most of the genome does not encode for genes, it also contain so called “gene deserts” that we have so far not understood the function of. These deserts are long stretches of DNA that do not seem to contain any genes that we know of. There are only about 23,000 genes in the human genome and only ~2% of the genome is used for coding

sequences [1, 2]. Some of the former gene deserts have been found to contain long range regulatory elements for genes, microRNAs regulating protein expression or long-coding RNAs (lincRNAs) which lack any known function.

Population admixture, migrations, genetic drift and selection formed the human genetic landscape of today and with molecular techniques one can study the similarities and differences between populations and investigate how genetically related different populations are [3].

In an individual, by using any of a number of genotyping techniques, one can establish exactly which base (A, C, G or T) the individual has at a specific base pair position in his or her genome. The combination of these two alleles on the two homologous chromosomes is the individuals' genotype at that position. If the same base is present on both chromosomes the individual is said to be homozygous at that position. If they differ the individual is heterozygous.

In summary, a human body contains a lot of variation, but these variations rarely have any impact on cell function. When drawing a blood sample, it contains many cells thus giving a mix of DNA from them all. When genotyping or sequencing the sample, the bulk DNA sequence which is the "original" setup inherited from the parents is obtained. One base change here and there not inherited from the parents is not detectable with the common standard methods of today.

1.1.1 Definitions of concepts in genetics

1.1.1.1 Knowing the phase, owning the phase

The human chromosome is a stretch of DNA containing a part of the human genetic code. The human species has 23 chromosome pairs that contain all the human genes (the X and Y chromosomes count as one chromosome pair but they contain different genes to some extent). Most cells have two sets of the genetic code and every gene is present in two copies, although sometimes only one of them is active, sometimes both of them and sometimes none (depends e.g. on tissue, activation and silencing, chromatin remodeling and imprinting). The chromosome is made up by base pairs that are connected with each other in a long chain (syntenic bases), and the bases come in a specific order but with variations which the 1,000 Genomes Project is trying to establish (<http://www.1000genomes.org/>). Having information on the consecutive sequence of bases on the same chromosome is called "knowing the phase" and is a very desirable knowledge (Example A and B). In case control cohorts one can only estimate the phase and until we get single strand sequencing, which is under development, we are unable to separate the two homologous chromosomes within one individual.

Example A. When sequencing the individual shown below (individual a/b) one would get the genotype G/C on position 8 and T/A on position 18, and all other bases would be homozygous. One cannot establish if G at position 8 and T at position 18 are on the same chromosome or if G and A at these positions are on the same chromosome with common methods.

Chromosome 1a: ACTGGTTGAAGTCAGGGTTCCTGAGTC

Chromosome 1b: ACTGGTTCAAGTCAGGGACCTGAGTC

Sequencing gives: ACTGGTT(G/C)AAGTCAGGG(T/A)CCTGAGTC

Example B. A new variation appears on position 23 where individual a/b was homozygous for A when sequencing another individual. Variation at position 18 is absent in individual c/d.

Chromosome 1c: ACTGGTTGAAGTCAGGGTCCCTGAGTC

Chromosome 1d: ACTGGTTCAAGTCAGGGTCCCTGGGTC

Sequencing gives: ACTGGTT(G/C)AAGTCAGGGTCCCTG(A/G)GTC

1.1.1.2 Haplotypes

In example A and B, there are three haplotypes present. A haplotype is a defined stretch of a chromosome. In this case there is information on each base but you can shorten the haplotypes to only the bases that vary.

Chromosome 1a: ACTGGTTGAAGTCAGGGTTCCTGAGTC or short: GTA

Chromosome 1b: ACTGGTTCAAGTCAGGGACCTGAGTC or short: CAA

Chromosome 1d: ACTGGTTCAAGTCAGGGTTCCTGGGTC or short: CTG

In families one can use the pedigrees to establish what the haplotypes in the children must look like, especially when multi allelic markers are present. In case-control settings there are several algorithms available for estimation of haplotypes such as the EM-algorithm [4] implemented in the software Haploview or the quasi-Newton algorithm [5] implemented in the software UNPHASED.

1.1.1.3 Linkage equilibrium and disequilibrium

During the human evolution there have been recombinations between chromosomes that cause the variations in the genome to mix. Recombination occurs when a germ line chromosome is divided and joined with another divided chromosome. Most often it is the homologous chromosome inherited from the other parent but sometimes the chromosome is joined with a different type of chromosome and this can potentially rupture gene functions or induce trisomy (cause of e.g. Down syndrome).

If the alleles of two different SNPs are found together on the same haplotype (i.e. on the same chromosome) as often as expected (based on their allele frequencies) those alleles are said to be in linkage equilibrium. If, however, the alleles are found either more often or more seldom than expected, linkage disequilibrium (LD) is present for these markers. New mutations occur on the ancestral haplotype and are in LD with the other variants on this haplotype. SNPs close to each other are more likely to be in LD compared to SNPs further apart, although the amount of variability makes it impossible to assume that neighbor SNPs are in LD. Variants may also be located in a conserved region (also called conserved synteny) where recombination does not occur often due to, for example, steric hindrance or protecting factors inhibiting recombination. The degree of LD can be calculated based on the frequency of markers and genotypes and there are two measurements, D' and r^2 , that I have used in the thesis papers. They are calculated with the help of the disequilibrium coefficient D [6].

$$D = \Pr(AB) - \Pr(A) * \Pr(B)$$

Where $\Pr(AB)$ is the proportion of AB haplotypes, that is the frequency of AB haplotypes. $\Pr(A)$ is the frequency of the A allele on SNP1 and $\Pr(B)$ is the frequency of the B allele on SNP2. The disequilibrium coefficient can take values between -0.25 and +0.25, 0 meaning that the proportion of AB haplotypes equals the expected value derived from the allele frequencies.

With the help of D one can calculate the total correlation between two markers in the genome.

$$r^2 = D / \{ \Pr(A)[1-\Pr(A)] * \Pr(B)[1-\Pr(B)] \}$$

r^2 can take values between 0 and 1, where 0 means no association correlation between the markers, and 1 means total association correlation.

Lewontin's D' is derived from D and is corrected for the maximum value that D can take given the allele frequencies. D' can vary between -1 and 1.

1.2 PATIENT AND CONTROL COHORTS

Several cohorts have been used in this thesis, collected in Sweden, Norway and Canada. There is a balance in reducing power (power is the probability of rejecting a null hypothesis when it is false) of a study by mixing populations and introduce heterogeneity, but at the same time increase power by increasing the numbers of individuals studied. To avoid reduction of power, the populations should be as similar as possible. Even though I did not detect any significant differences between the Norwegian samples and the Swedish samples I chose to include "country of origin" in the logistic regression models in paper I and II just in case some hidden pattern could be present underneath the allele frequency distribution.

In paper III where Canadian, Norwegian and Swedish samples were included, we only analyzed carriage of one human leukocyte antigen (HLA) allele and the frequency was similar between the countries.

1.2.1 Study design

In the four articles included in this thesis, case control design was used in three and a case-only design in the fourth.

1.2.1.1 Cohort study design, why do we not use this?

Case control design is commonly used when studying diseases that are not so common in the population. The "opposite" of case control studies are cohort studies where you take a random sample of the population and study the number of individuals that develop disease over some period of time [7]. This is not very effective in a rare disease since large samples are required to reach a sufficient number of cases. It can answer several interesting questions such as how many individuals in the population develop disease and at which rate this is happening. But when studying how genetics is involved in the development of disease the number of cases in such a study will be too few to get a reliable result even though the prevalence of MS in Sweden is 188.9 per 100,000 individuals [8]. It would require a starting cohort of around 20 million people

to obtain 1000 MS cases after 1 year or 4 million people followed for 5 years since the incidence in Sweden is about 5 MS cases /100,000 individuals per year.

1.2.1.2 Case control design

A more effective approach when wanting to study genetic influence in a less common disease such as MS is the case control study design where patients are ascertained through clinics to participate in the study. To be able to study the effect a certain genetic variant has on susceptibility of disease, a comparison between patients and individuals that do not have disease must be made. Hence, a group of healthy individuals are collected and these are called the “controls”.

1.2.1.3 Case only design

Sometimes the variable of interest is not measurable in controls, e.g. disease progression or response to treatment. The fourth article in this thesis have a case only design where we studied the outcome of neutralizing antibodies (NAbs) in interferon beta (IFN- β) treated patients and if HLA variants had any effect on NAb development.

1.3 METHODS OF ANALYZING THE DATA

1.3.1 Odds Ratio

Odds Ratio (OR) is a measurement where the odds of being exposed for a variable, for example having a specific genetic variant, among patients is divided by the odds of being exposed to the same genetic variant in the healthy individuals. The formula looks like this, (*DRB1*15* is a genetic risk variant for MS):

OR= Odds among patients/Odds among healthy

Odds among patients = $N_{\text{case}}(\text{DRB1*15 positive}) / N_{\text{case}}(\text{DRB1*15 negative})$

Odds among healthy = $N_{\text{ctrl}}(\text{DRB1*15 positive}) / N_{\text{ctrl}}(\text{DRB1*15 negative})$

Odds ratio is used to describe how the odds of exposure for patients is related to the odds of exposure within the healthy individuals. An OR of 1 means that the odds of exposure among patients are equal to the odds of exposure among healthy individuals, and this is interpreted as that the exposure have no effect on disease susceptibility. A higher value indicates a higher odds among patients (risk variable) and a value lower than 1 indicate a decrease of exposure in patients compared to healthy individuals (protective variable).

1.3.2 Null hypothesis, power and significance of a test statistic

A scientific hypothesis is a proposed explanation to a certain observation, a relationship between variables that results in the outcome. Scientific hypotheses are translated into statistical hypotheses that can be tested. A null hypothesis normally states that there is no relationship between the variable and the observed outcome. Based on acceptance or rejection of the null hypothesis the corresponding scientific hypothesis is either supported or not.

Power is the probability to reject a null hypothesis when it is truly false, i.e. the sensitivity of a statistical test. Power assessment is most effective when designing a

study to see how many patients and controls should be collected in order to evaluate a previously reported or a biologically relevant effect of a variable. It can also be used to estimate a variables maximum effect size that can be detected in a sample of a certain size.

A P-value is a measurement of how likely it is that the observed result or a more extreme one would occur by chance if the null hypothesis is true. The significance level (α) is the cut-off chosen for the P-value in order to denote the result “statistically significant”. A significance level of 0.05 is commonly used and a statistically significant result would then imply 5% risk or less of committing a type I error i.e. of rejecting a null hypothesis that in fact is true.

2 HUMAN LEUKOCYTE ANTIGEN

2.1 HLA MOLECULES

Within all vertebrates there is a genomic region encoding immune molecules called the major histocompatibility complex (MHC). Many of the immune molecules encoded in the MHC have a central role in the immune system when it comes to antigen presentation. In humans the antigen presenting molecules are called human leukocyte antigen molecules and can display a peptide within their binding clefts that other immune cells can either recognize or ignore depending on their education in the thymus during their maturation. One can say that the HLA molecules act as a store window that you (resembling a T-cell) pass on the street. If it presents something that does not suit you, or that you are not interested in, you just pass on by, and thereby ignoring it (self-molecules). If it presents something you recognize as not appropriate or something unusual that makes you have a closer look you will decide if you should react by your favorite mechanism, for example protesting on the street (alerting the rest of the immune system, T-helper cell) or kill the store owner (cytotoxic T-cell or NK-T-cell) or just ignore it (tolerance). However, sometimes the cells protest even if there is nothing a normal immune cell would call strange in the store window, that's when autoimmunity is born. T-cells start to react towards self-molecules that should not evoke such a reaction. There are several mechanisms behind this type of misbehavior and one model is epitope spreading. If using my former example, epitope spreading can easily be explained in the following way: suppose there is a new large expansion of some store that show a lot of naked skin (antigenic peptides) in their windows, the crowd (immune system) gets upset and starts to react by forcing the stores to shut down and people (immune cells) participate in raids on the streets (one could only imagine what happens within those four walls!). After a while most of those stores are gone, out of business (infection is over and taken care of), but the crowd (T-cells) is still upset and reacts also on stores that have mannequins in lingerie because it makes them remember the almost naked people they just vanquished from the streets. They can also react to all the debris left from the riots, the tissue damage itself can activate immune response. Although the mannequins were tolerated and considered harmless before the invasion they are now considered harmful for society (non-self). This reaction can be triggered long after the first infection is gone as long as there are T-cell clones from last infection left in the system. This, so called long-term memory of antigens is also one of the strengths within the immune system. The next time the same pathogen peptide appears on the HLA molecule, the cells will react much faster than last time.

The human MHC gene region, i.e. the HLA gene complex, is located on the short arm of chromosome 6 and spans over 7.6 Mb (Megabases). The region is divided into three compartments called class I, class II and class III (Figure 1). I will not discuss class III much in this thesis since it does not contain any of the classical HLA genes in focus of these studies. The class III gene region is situated in-between class I and class II and harbors many genes of which a majority are important in immune response.

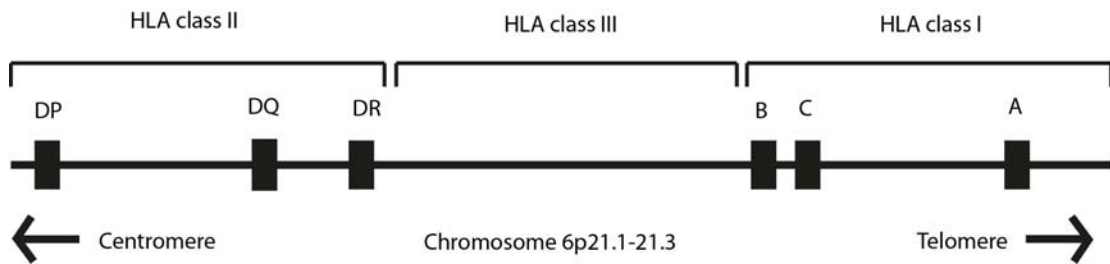


Figure 1. Schematic map over the HLA region on chromosome 6. The centromere is located towards the middle of the chromosome and the telomere at the ends. HLA gene names are shortened in the figure.

Both class I and class II HLA molecules present peptides to immune cells from the cell surface but there are some differences between them that will be briefly discussed in the next section.

2.1.1 Class I molecules

Each class I molecule consists of two peptide chains of which one is the invariant β -2-microglobulin (β_2m), also called the light chain, encoded on chromosome 15. The other peptide chain, the heavy chain, is polymorphic and encoded by the genes *HLA-A*, *HLA-B* and *HLA-C*. Class I genes are expressed in all nucleated cell types. The expression of *HLA-C* is not as abundant as that of *HLA-A* or *HLA-B*. It has been described that the *HLA-C* heavy chain poorly associates with β_2m [9] resulting in low expression on the cell surface.

The HLA class I molecules are assembled by association of the heavy and light chains and stabilized by the binding of a peptide in the binding cleft formed within the heavy chain. When there is not yet a peptide in the peptide binding groove, the molecule is chaperoned by other proteins (called the peptide loading complex, PLC) until a suitable peptide can bind [10]. The HLA class I molecule is loaded in the endoplasmic reticulum with a 8-9 amino acids long peptide and transported to the surface. These peptides are usually formed by degradation of endogenous molecules by the proteasome. Most peptides presented in this way are self-peptides, parts of degraded proteins produced within the cell, but not all peptides found on HLA class I are coded for within the genome. Proteases can cut and paste peptides into variants not normally present in a cell [11]. The class I molecules are especially important when a virus invades a cell and starts replicating within it. Then HLA class I molecules will present virus peptides (non-self) on the surface of the infected cell and alert the immune system to react.

There are some differences between the peptide loading of *HLA-A*, *-B* and *-C*. The PLC stabilizing the HLA class I molecules include a molecule called tapasin which can trim peptides about to bind the HLA class I complexes. This process enhances the variability of the peptides yielded by the proteasomes. *HLA-A* and *HLA-C* bind the PLC quite nicely while *HLA-B* does not [12]. *HLA-B* molecules can instead be loaded with peptides directly in the lumen of the endoplasmic reticulum, and the HLA peptide complex is therefore transported to the surface faster than *HLA-A* and *-C*. Most *HLA-B* molecules are loaded with peptides, while only 30-70% of the *HLA-A* and *HLA-C* molecules are loaded [9].

2.1.2 Class II molecules

The HLA class II molecules consist of two amino acid chains, one α -chain and one β -chain of equal size. Both chains can be polymorphic but the β -chains are by far the most polymorphic and are encoded by any the three classical genes *HLA-DRB1*, *HLA-DQB1* and *HLA-DPB1*. The α -chains are encoded by the *HLA-DRA*, *HLA-DQA1* and *HLA-DPA1* genes respectively. On some haplotypes there are extra *HLA-DRB* genes thought to originate from duplication of the *DRB1* gene. These are called *DRB3*, *DRB4* and *DRB5* and are also expressed but at a lower level than *DRB1*. The chains pair up in the endoplasmic reticulum lumen and associate with a peptide binding cleft molecule called the invariant chain which prevents binding of self-peptides that are restricted to HLA class I molecules. The molecules are then transported to some late endosomal compartments, containing peptides derived from the endosomal pathways. The invariant chain is degraded and the HLA molecule binds a 14-20 amino acids long peptide. The HLA class II molecules have a binding cleft that is open in both ends which facilitate binding of longer peptides than HLA class I. The molecules are then further transported to the surface for antigen presentation to $CD4^+$ T-cells. The HLA class II molecules present mainly peptides derived from the surrounding environment of the cell. Extracellular components engulfed by the cell are degraded via endosomes and lysosomes. The peptides generated in that process will bind to HLA class II when the invariant chain is degraded. HLA class II molecules are normally only expressed on antigen presenting cells (APCs) but other cells can induce expression if exposed to IFN- γ .

2.1.3 Cross presentation

The restrictions of peptide binding described above are usually what the bulk HLA molecules present. There are pathways that introduce foreign peptides on class I molecules and self-peptides on class II molecules. One pathway that has been in focus lately in autoimmunity is autophagy, a mechanism for regulation of protein turnover by engulfment of cytoplasm. The content of the autophagosome is then degraded via the lysosomal pathway resulting in self-peptide presentation on class II molecules [13]. Autophagy is heavily studied and considered as a mechanism of principal importance in Crohn's disease [14].

2.2 GENETIC VARIABILITY WITHIN HLA

There are structural and genetic similarities between the HLA molecules that are thought to originate from duplication of genes during evolution. Variability of the HLA has helped the human species to survive through times of selective pressure by, for example, pathogens. With a wide range of variants in a population, the probability of some individuals to survive an infection of a fearsome pathogen increases and thereby the species survives at the expense of those carrying variants not so beneficial. For every generation, variability changes within a population and to quote something I heard in a seminar –“Nature doesn’t care about the individual, it only cares about the species” summarizes this quite well. The HLA gene region on a chromosome can be seen as a patchwork of different variable regions that are highly polymorphic between individuals. However, certain combinations are more often seen than others within a

population. One such example is the ancestral B8 haplotype that has been intensely studied. It contains *HLA-A*01*, *HLA-B*08* and *HLA-DRB1*03* and is the most common haplotype in Caucasians. The naming of HLA gene variants is complicated and the HLA nomenclature has been revised several times during the years. An example can be seen in figure 2.

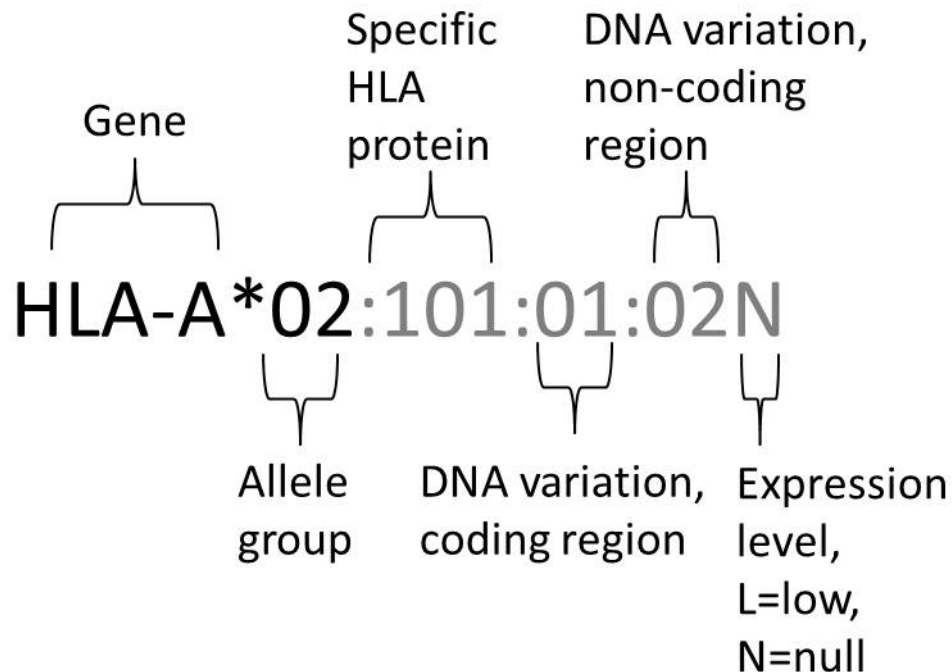


Figure 2. This is an example of the naming of an allele. Picture adapted from www.hla.alleles.org. This thesis focuses on the allele groups, black portion.

The patchwork of combinations of these chromosomal regions put together, forming the HLA haplotypes we see today, is a result of many generations of recombination, selection, duplications and deletions of genes. Some haplotypes are more common than others and some are even new. It is a characteristic of the HLA gene region that makes it problematic to analyze its effect on disease susceptibility. Grouping the genetic material into variants that encode proteins that share physical properties in a similar way or a subgroup within those, is a way to reduce heterogeneity between the individuals. But in doing so, one usually have a hypothesis of on what level of HLA resolution the susceptibility to the disease is present.

There are many alleles in each classical gene and the most diverse of them is *HLA-B* with a remarkable 2,338 alleles registered in the IMGT/HLA database (<http://hla.alleles.org>, accessed 6 April 2012). These alleles encode 1,795 proteins, meaning that almost 77% of the genetic variants encode unique proteins (Table 1).

Table 1. Number of alleles and protein registered in the IMGT/HLA database at the date of access.

Classical gene	Alleles	Proteins	Ratio P/A
HLA-A	1757	1290	0.73
HLA-B	2338	1795	0.77
HLA-C	1304	946	0.73
HLA-DRA	7	2	0.29
HLA-DRB1	1166	873	0.75
HLA-DQA1	47	29	0.62
HLA-DQB1	162	113	0.70
HLA-DPA1	33	16	0.48
HLA-DPB1	152	131	0.86

P/A ratio = number of proteins / number of alleles

All this variation is a source of endless information but makes it hard to determine which exact variant an individual have. There are today several methods that can be used and in my studies many of them are represented. I have myself used the Olerup kit with sequence specific primers (Olerup SSPTM HLA low resolution Kit [15]) that detects the serological type, i.e. at the two digit level (Figure 2). This level of HLA resolution corresponds to groups of proteins that respond similarly to specific stimuli. Genotyping of *HLA-B* was done by personnel at the department of Clinical Immunology and Transfusion Medicine at Karolinska Sjukhuset Huddinge which had a long experience of HLA genotyping with a Luminex based reverse PCR sequence specific oligonucleotide method (LabType® SSO from One Lambda, Inc., Canoga Park, USA). This method detects PCR amplified sequences with beads and returns groups of alleles positive for that sequence. I reduced those to two digit level for our analysis.

With large genome wide association studies (GWASs), with large amount of SNPs genotyped throughout the HLA region it is possible to impute alleles based on LD between the SNPs. This method is still under development as more and more populations are genotyped and the results so far is that imputation is fairly good, excellent for some alleles but not so good for others. This is probably due to that the SNP chips used only can design probes for some SNPs and not others and that not so common alleles are avoided.

3 HLA IN COMPLEX DISEASES

MS is a complex disease, or a multifactorial disease, meaning that it is not caused by one single factor but rather a combination of factors. An even more complicating aspect is that these factors may not be the same for all patients. It probably depends on underlying interactions between factors if disease is developed or not. It is for example not enough to smoke to get lung cancer even if it in itself is a strong risk factor. Many people smoke their whole life and never develop lung cancer and some people get lung cancer even though they have never smoked. The risk factor has reduced penetrance meaning that the probability of getting disease if positive for the risk factor is not 100%. An individual has a genetic make-up that makes it unique. As a result of genetics and surrounding factors the individual will take certain steps in life that expose the individual for all kinds of different environments that affect susceptibility for life time events such as accidents or diseases (both chronic and temporary). Some will develop this or that disease during their lifetime, some will not. Those factors that we manage to pinpoint as risk factors for disease are not exclusive for the patients; otherwise it would be a “monofactorial” disease. These factors are to some extent also present in controls and therefore some other, unknown contributing effect induces disease in combination with the known risk factor.

3.1 HLA IN AUTOIMMUNE DISEASES

Several diseases with features common to MS are associated with certain HLA alleles, especially autoimmune diseases. The larger genome wide association studies have found several genes that seem to be important in many diseases pointing toward key factors that could tell us more about different disease pathways. However, most often does one allele increase risk of one disease and the other allele increase risk for another disease. This scenario fits rather well with the notion that the genetic landscape one carries together with the environmental factors increase the risk for some diseases and decreases it for others. HLA genes are such genes with different risk for disease depending on variant, but what role HLA play in these diseases is not always known.

3.1.1 Type 1 Diabetes

Type 1 diabetes (T1D) is an autoimmune disease where the insulin producing β cells in the pancreas are destroyed in a process that starts already in childhood. HLA class II genes account for up to 50% of the genetic risk of T1D, as reviewed in [16]. The class II variants responsible for this association are two positively associated haplotypes, one being *DRB1*03,DQA1*05:01,DQB1*02:01* shortly called DR3 and the other one is *DRB1*04,DQA1*03:01,DQB1*03:02* shortly called DR4. These are common haplotypes in Caucasians, but up to 95% of all T1D cases carry at least one of them. There is one negatively associated haplotype, the MS risk haplotype, *DQA1*01:02,DQB1*06:02* which is dominantly protective in T1D. The rest of the haplotypes confer a spectrum of risks varying from protective to mildly increasing the risk of T1D.

3.1.2 Rheumatoid arthritis

Rheumatoid arthritis (RA) is an autoimmune disease where chronic inflammation of smaller joints is the main symptom. About 1% of the world wide adult population is affected. The HLA is the most important genetic risk factor in RA and accounts for up to 50% of all the genetic risk for RA (reviewed in [17]). There is a “shared epitope” hypothesis applied to HLA in the RA field, where a conserved amino acid sequence can be found within several allele groups. These shared epitopes can be found in *HLA-DRB1*04* and *DRB1*01* and it is only associated with ACPA (anti-citrullinated protein antibodies) positive RA, thereby distinguishing a genetic difference between ACPA positive and negative RA. There is a hypothesis that the shared epitope HLA molecules are involved in shaping the T-cells to become autoreactive. In ACPA positive RA, protection is associated with *DRB1*13:01*. The ACPA negative RA is instead associated with *DRB1*03* that is not part of the shared epitope allele groups and there seem to be no protection from *DRB1* alleles for this type of RA.

3.1.3 Systemic lupus erythematosus

Systemic lupus erythematosus (SLE) is an autoimmune disease where symptoms can vary and diagnosis is made on basis of a set of criteria. The main phenotype is that the patient is positive for autoantibodies directed against nuclear antigens. In SLE, the main associations to HLA are found with three haplotypes, *DRB1*15:01,DQB1*06:02, DRB1*08:01,DQB1*04:02* and *DRB1*03:01, DQB1*02:01*, as reviewed [18]. However, there are also genes in the HLA class III region that are important for SLE susceptibility. Among those are complement genes involved in innate immunity.

3.1.4 Celiac disease

Celiac disease is an autoimmune disease that is induced by eating gluten and can therefore also be reversed by stopping the intake of this protein. The disease has a strong association to *HLA-DQA1*05, DQB1*02* and *DQA1*03,DQB1*03:02*, the same as the major risk haplotypes for T1D. These two haplotypes are the main risk variants and those that do not carry either of these two, carry haplotypes where *DQA1*05* or *DQB1*02* is present. These HLA allele groups are necessary for developing disease but not always sufficient [19]. It is thought that very stable gluten peptide binding to these HLA molecules is responsible for the strong activation of the immune system. The perfect matching shape of the peptides is enhanced by the activity of transglutaminase 2 enzyme.

3.2 MULTIPLE SCLEROSIS

Multiple sclerosis (MS) is a chronic demyelinating and neurodegenerative disease of the central nervous system (CNS) that has a strong inflammatory component. It is diagnosed by a wide set of criteria of which some combinations have to be fulfilled. The criteria have been revised during the years, such as the Poser or McDonald criteria [20, 21]. Hallmark investigational findings carried by many patients are presence of IgG antibody production (called oligoclonal bands) in the cerebrospinal fluid (CSF) not seen in plasma and inflammatory lesions in the brain detected by magnetic resonance imaging (MRI) separated in time and location. Examples of symptoms in MS patients

are fatigue, numbness, muscle weakness, pain, cognitive impairment and blurry vision [22]. MS is divided into two courses and about 5-10 % of patients follows the primary progressive course (PPMS), gradually getting worse but do not really have any bouts of symptoms. Relapsing remitting course (RRMS), also called bout onset MS, is the most common form and patients get bouts of symptoms but recover back to normal function in between. As time goes by the ability to fully recover between the bouts is reduced and the patients gradually enter the secondary progressive phase of bout onset MS (SPMS). In this phase the patients gradually worsen and the bouts are more or less absent as in primary progressive course. Patients with bout onset MS can have very few bouts or very many and every number in between, it can vary through the years and the length of the relapses differs from patient to patient and from time to time. MS is more common in women than in men while men more often develop the PPMS subtype. A geographical gradient describes a higher frequency of MS in areas closer to the poles than to the equator, a finding that has stimulated hypotheses on the effects of sunlight, vitamin D and ancestral genetic heritage from the Vikings.

In monozygotic twin pairs where one individual is affected with MS, only 1/5 of the co-twins are also affected. This is usually a way to describe that not only genetics is part of disease susceptibility since monozygotic twins are considered genetically identical. On the other hand, the risk for the healthy twin to develop disease is much greater than for an individual in the general population with no relatives diagnosed with MS (lifetime risk 1/500), thereby indicating that genetic predisposition also is important.

Recombinant interferon beta (IFN β) is a first-line therapy for RRMS. The therapeutic effect of IFN β in MS has been investigated in a number of trials, which have shown that IFN β reduces the number of clinical relapses and new lesions, as well as the accumulation of disability over time [23-25]. The understanding of the mechanisms behind the disease-modifying effect of IFN β in MS is still incomplete.

3.2.1 HLA in multiple sclerosis

The strongest genetic risk factor for MS found so far is the HLA class II haplotype *DRB1*15:01,DRB5*01:01,DQA1*01:02,DQB1*06:02* [26] efficiently tagged already in the 1970-ties by association to class I alleles [27, 28]. This haplotype is present in about 60% of the patients and 30% of the controls in our cohort [29] and association to increased risk of MS is found in most populations worldwide. Today the negative association to MS of *HLA-A*02* is also established [29-32]. The function of HLA in MS susceptibility is not clear as it is also debated if the inflammation of the brain precedes demyelination thereby driving disease, or if it is a result primarily from demyelination and that the immune components are trying to clean up the tissue damage. Lately, in part due to all the GWAS results indicating roles of immune genes, increased evidence point in favor of the former hypothesis [33]. At the same time, new studies of neurodegeneration in the brain increase the evidence for the later hypothesis, as reviewed in [34, 35]. Thereby both hypotheses gain more evidence and research around the early diagnosed MS cases might help to resolve this question.

What has been observed is that CD4⁺ and CD8⁺ T-cells seem to have a central role in pathogenesis. It is thought that a lesion forms when, for some unknown reason, T-cells

start attacking the myelin producing oligodendrocytes in the brain. Myelin basic protein (MBP) and myelin oligodendrocyte protein (MOG) are two components of the myelin-derived isolating sheaths around the axons in the CNS, and the proteins are released when the myelin sheath is disrupted. Without the myelin sheath, the axons are no longer protected from surrounding influx of ions and the conduction of neuronal impulses is heavily reduced or even eliminated. The blood brain barrier close to these lesions is often disrupted making it easier for immune cells to enter that normally would not access the CSN. The destruction of myelin sheaths, axons and also nerve cells results in the loading of HLA molecules on surrounding APCs with a smorgasbord of self-peptides presented on both class I and class II. All the cytokines and other released stimulatory molecules activate upregulation of co-stimulatory molecules on the APC surface and this is where the hypothesis of epitope sharing comes in. Some T-cells recognize peptides that closely resemble for example myelin peptides. One such T-cell has previously been activated by some infection, recognizing the peptide as a “threat” because the APC also display the co-stimulatory molecules that trigger an “alarm”-signal in the T-cell. When this cell then finds an almost identical peptide generated by degradation of myelin, the recognition and the inflammatory milieu in the lesion activates the T-cell again and it starts to proliferate and generate many more cells of the same kind and autoimmunity against the myelin derived peptide is born. This “epitope spreading” is a part of a larger concept of molecular mimicry which is an example of how lesions can arise and prevail [36], but there are some evidence that support it. Many viruses have been investigated in MS but few reports have been confirmed. There is increasing evidence for a role of previous exposure to Epstein-Barr virus (EBV) preceding the MS diagnosis and a potential link to HLA [37, 38]. EBV derived peptides could be the first step in the epitope spreading hypothesis. MOG specific T-cells have been isolated in MS patients and are more frequent in patients compared to healthy controls [39]. MOG peptides can potentially activate these T-cells.

Studies in animal models have frequently been reporting evidence of how MHC might be involved in regulation of experimental autoimmune encephalomyelitis (EAE, the MS rodent model). EAE can be induced in susceptible strains by injection of MOG, MBP or proteolipid protein (PLP) and the symptoms are similar to RRMS with relapses and remissions. Transgenic mice with human class II alleles and no mouse MHC have been used to investigate the role HLA have in MS immunopathogenesis. It seems like DR molecules are required for EAE susceptibility while DQ molecules act as risk modifiers in mouse models, reviewed in [40]. This hypothesis goes hand in hand with reports of a Canadian family material where MS family trios have been studied for modifying effects of *DRB1*15* [41]. It was however not possible to detect such effects due to strong LD in the large GWAS done on 9,772 MS cases and 17,376 controls [32].

4 MY STUDIES

4.1 PAPER I AND II

The idea to the first paper in this thesis was already in discussion within the group when I started my PhD in Huddinge. This was my first project and I engaged in genotyping *HLA-C* in a cohort where genotypes were already known for *HLA-DRB1* and *HLA-A*, as part of a project that my supervisor Boel Brynedal published around the same time I started in the group [29]. The research group had a good relationship with Hanne Harbo's MS genetics group in Norway and it was decided to make a joint project in order to increase power. The aim was to investigate if further signals in HLA class I was present in MS and/or if the *HLA-A*02* signal presented by Boel was a proxy for some other signal that could be picked up by an allele group in *HLA-C* which is located in between *HLA-DRB1* and *HLA-A* on chromosome 6 (Figure 1). At that time there were doubts against signals in HLA class I in MS genetics and especially if presented with "complex" statistical tests like logistic regression which was met with hesitancy by several reviewers. Originally, I didn't understand anything Boel explained regarding her analysis and after showing me repeatedly, she must have been frustrated.

4.1.1 The concept of LD within my data

When typing DNA all days long, running electrophoresis gels and taking pictures at the UV-table, I started to get a grip of the complexity of the HLA, both by reading papers that were published, the genotyping data I produced and by all the discussions we had within the group. The concept of LD started to sink in when I saw that it was very common in my dataset to be *DRB1*15* positive and *C*07* positive but I could not see that clear relationship when looking at *HLA-A*. I started to form my picture of HLA, diplotypes, extended haplotypes and evolutionary conserved haplotypes and realized how much more complex it actually was. Was there anything that was not dependent on anything else?

4.1.2 Scooped?

While I was genotyping, several papers, reported class I associations to MS. One of the first ones was done by the International multiple sclerosis genetics consortium (IMSGC) and the authors reported an association to *HLA-C*05* [42]. Their study had the same aim as mine but used a mixture of SNP data, microsatellite data and low resolution genotyping data in a cohort of British family trios and sporadic cases as well as US family trios. The primary feeling was "scooped" but after some discussions we realized that there were some important differences between the studies. They had to infer their allele groups from different types of data such as micro satellites, SNPs and normal HLA genotyping; we had HLA low resolution genotyping from only one source in our cohort, giving our study an advantage. Our cohort was almost as large as theirs; we thought that if we would see the same results, our paper would be a nice replication. We also wanted to study our cohort with another method than what would be used in a replication analysis. Those authors practically gave me the instructions in how to perform a statistical analysis of HLA association in that paper. Also, we knew already at this stage that we wanted to genotype *HLA-B* due to recent reports of negative

associations to *B*44*, so this would be a perfect practice for me. It was therefore decided to continue the study.

4.1.3 Several reports but only one main point

While I was analyzing the data and writing the paper several other reports of secondary signals within the class I region was published. One of those suggested an association in the MOG gene located telomeric of *HLA-A* [43]. In that study, no information on classical HLA class I genes was available and a connection to *A*02* or *C*05* could not be studied. Another study used SNP mapping and candidate gene mapping in analyzing the connection between MOG and *HLA-A* in Tasmanians [44]. This study did not have *HLA-C* data and a possible effect of *C*05* could not be studied. In addition to these other reports an article published by the IMAGEN group was in press before our paper was accepted [45] showing a more complex analysis based on SNP data. The authors discussed a secondary signal best explained by *B*44:02* that emerged when conditioning on a SNP in LD with *DRB1*15*. Cree and colleagues could not rule out *HLA-A* since there were secondary signals also around that locus although not as strong as in the *HLA-B/-C* locus. Also, a report of association to *HLA-B*44* was found in a meta-analysis of several GWAS scans and the signal was independent of *DRB1*15* [46].

These papers pointed in different directions but they all said one thing: There is something hiding in the class I region and it might be independent of *A*02* or even an association superior of *A*02*. The most fundamental problem, apart from the lack of *HLA-B* data, was the statistical approach. How does one take into account the extensive LD but still highly polymorphic genes without losing power and using overly complicated models? A typical way of analyzing HLA was to stratify the cohort on the most associated marker and exclude those individuals from further analysis [47]. This approach did not seem right, although I couldn't put my finger on why. It didn't make sense to ignore the other chromosome in those individuals that carried the most associated allele group (in our case *DRB1*15*). What if the carriage of some other allele did affect the risk of *DRB1*15*? There were studies that indicated that that might be the case [48, 49]. Even though our cohort was not powerful enough to investigate that type of effect modifications did not mean that I could ignore it. Actually it could even mean that I possibly would draw the wrong conclusions from the results. Additionally, a stratification approach always reduces sample size, which decreases your power.

4.1.4 Recursive partitioning didn't work

In an article from the Type 1 Diabetes Genetics Consortium (T1DGC) the authors used recursive partitioning to establish risk haplotypes between *DRB1* and *DQB1* [50], and the authors later corrected for a haplotype in regression analysis when searching for class I signals independent from class II in T1D. This was a way of reducing the influence of HLA class II heterogeneity when searching for independent signals in HLA class I. I thought that maybe this could be used within HLA class I to distinguish which allele groups had an independent effect of *A*02* and *DRB1*15*. The method splits the material on the variable that best separates cases from controls. It continues to split the cohort in a tree like fashion until the defined maximum number of nodes is reached or the purity of the nodes is no longer significantly better. The first split

separated the cohort into one node with *DRB1*15* positive and one with *DRB1*15* negative individuals. The purity, i.e. separation of cases from controls, of the nodes was significantly better than in the starting material. The second split divided both nodes on *A*02* carriage, indicating that the effect of *A*02* was independent of *DRB1*15*. The purity of the groups was better, but the improvement was not on the same scale as the split on *DRB1*15*. An analysis of purity and complexity of the trees indicated that we did not have enough power to study more splits than the first two, the gain of purity did not compensate for the increasing complexity of the tree. We could not use this method to study effects other than *DRB1*15* and *A*02* that we already knew of.

4.1.5 Mimicking the IMSGC paper

To be able to compare the results to the already published paper from IMSGC we had to analyze our data in the same manner. The method removed the largest risk allele group and thereafter the largest risk allele group in the remaining dataset. First the IMSGC authors excluded *DRB1*15*, then *DRB1*03* and after that *DRB1*01:03* and within the individuals negative for all three alleles, only *HLA-C* was globally significant and within that gene the strongest signal was *HLA-C*05* which was negatively associated to MS. Either we could do our analysis exactly the same way, excluding the exact same alleles or we could use the same method but apply it to our data and exclude the alleles that had further signals in our cohort, regardless of if it was a different allele group than in the IMSGC study. When excluding the exact allele groups (first *DRB1*15*, then *DRB1*03* and *DRB1*01:03*) as the IMSGC did in their study, both *HLA-DRB1*, *HLA-A* and *HLA-C* were globally significant and within each of them I had several signals from several allele groups. We decided to instead use the same method and apply it to our data. After exclusion of *DRB1*15*, the most significant gene was *HLA-C* and within that gene the *C*08* allele group was the most associated. After removal of the *C*08* positive individuals, weak signals could be found in *HLA-A* and this result was reported in our paper.

4.1.6 My way of thinking about logistic regression in paper I and II

In the paper that Boel published, about the same time I started in the group, a nested logistic regression was used to study the effects of *HLA-A* allele groups while having *DRB1* allele groups as covariates [29]. We decided to use this approach in paper I since the method provided a way to correct for already known risk alleles while still retaining information on the other chromosomes that would be excluded by a stratification approach. Since *HLA-C* is located in between *HLA-A* and *HLA-DRB1* we had the following main objectives: First, to try and map the effect of *A*02* and investigate how far the signal extend from *HLA-A*, and whether it was also in LD with something in *HLA-C*. The second objective was to investigate whether there were allele groups of *HLA-C* that were associated with MS independently of *A*02* and *DRB1*15*. For both these objectives we turned to logistic regression. The problem was, though, how to model the data to be able to draw conclusions in an effective way. First we wanted to see if the information on each gene could explain the outcome. When adding HLA genes to the regression model, the number of variables increases dramatically due to the highly polymorphic state of the genes and this decreases power. Even though this was a problem, we could see that all three genes added information that better explained the outcome. The fit was best for the *DRB1* gene and the next step was to correct for the

DRB1 gene and analyze if addition of any of the class I genes increased the fit of the model. This was true for both *HLA-A* and *HLA-C* but the overall fit of the *DRB1 + A* model was better than *DRB1 + C*. So the last model was to investigate if the data on *HLA-C* could increase the fit significantly over *DRB1 + A*, and this was indeed the case.

We could also see that individual allele groups from each gene were associated to MS in the final model. The alleles that were associated with MS were mostly *DRB1* alleles but both *A*02* and *C*08* was associated even after false discovery rate (FDR) adjustment. These were the same alleles that were found in the stratification analysis but we could not see the other *DRB1* allele group associations in that analysis. Hence, the conclusion might be that stratification is not the perfect way to study signals within the same locus but more powerful to detect a signal from a second locus.

4.1.6.1 *New insights in paper II*

When I got the *HLA-B* data and the first paper was submitted and more or less done I had learned a bit more about logistic regression and realized that the way we did it in the first paper was not optimal for the second paper. We compared the genes to each other with all allele groups added and in the final model there were almost 30 variables included. This is not at all optimal and probably not even possible to draw any conclusions from when I now had four genes to study. However, the number of variables in my study was a bit deceiving, actually none of them were independent from all others and how many variables were then acceptable? In the beginning of the analysis I used nested logistic regression as in the first paper but I had problems with interpretation of the results and what those actually meant. Was a larger model better just because it was more complex? Was it possible to exclude those alleles that were not significantly associated? How should we code the allele groups, was dosage more informative than carriage or did we assume too much with that model? The discussions within the group how to solve this went on every week on the Monday meetings, but I am not sure how far we came each week. One of the break points came one day from an unexpected source. A colleague was in Stockholm visiting our laboratory and as a statistician, immediately pointed out the obvious problem that the final model had too many variables. We had to reduce it somehow; there was no other choice in this case. That was all it took for turning the project around, we finally got some straight forward directions on how to pursue this project. I am very grateful for that input, and the funny thing was that we had discussed it but none of us actually knew what “too many variables” meant, where do you draw the line? It most certainly depends on the hypothesis and that’s when my brain started questioning that concept. What was our hypothesis exactly? Under which circumstances did this apply, and under what model? What was previously known that was relevant to use, what questions can logistic regression answer, what level of association is meaningful and in what setting should the questions be asked? The questions went on until a clearer picture of how I wanted to analyze the project emerged. A change in tactics was needed.

4.1.6.2 *Logistic regression reloaded*

While the genotyping of *HLA-B* was outsourced to save DNA and reduce my hours in the lab, new papers had come out that studied the signals in class I. Most of the papers reported independent effects of HLA class I allele groups on top of *DRB1*15* but no

overall conclusion was made. There were reports that the protective effect of *A*02* was enhanced by carriage of *C*05* and that *C*05* could be a risk factor if carried together with *DRB1*03* [51]. The authors only genotyped for the specific allele groups of *A*02* and *C*05* and could therefore not take any other allele group into consideration. A year later a follow up study was published where they had added data for presence or absence of *B*44* and *B*18* [52]. The main finding from this paper was again that *C*05* enhanced the protection of *A*02* and that the haplotype maintained the protection when correcting for *B*44* but the authors ignored the fact that *C*05* did not have any effect at all when correcting for *A*02* and *B*44*. *C*05* was not associated to MS when correcting for only *B*44*, and showed independent negative association when correcting for *A*02* in the family material they analyzed but not in their case control cohort. The allele groups seemed to be connected to MS susceptibility in some way that was not so easy to study. The protective association also seemed to be a bit different in different populations, in some the main effect came from *B*44*, in some from *A*02* and in some *C*05* modulated the protective effect. All this information taken together made me curious if they actually were one and the same signal but that it was masked between populations by allele group haplotype frequencies. This would also explain why the big IMSGC GWAS did not find any more signals in class I when correcting for *DRB1*15* and *A*02* [32]. Only *DRB1* alleles were reported to have residual effects in that study. That was how my hypothesis about a protective class I haplotype was born.

Going back to the scientific question, I didn't just want to study the previously reported associations of *A*02*, *C*05* and *B*44*. I had complete allele group information on all four genes in 1,784 patients and 1,660 healthy controls, genotyped specifically for the HLA genes and not inferred from SNP data, which was something no other group had. We felt obliged to investigate MS susceptibility in a more hypothesis free approach, at least to start with. So, what could I really use logistic regression for and should I exchange it for something else? I realized, when looking deeper into the concept of logistic regression, that it is mainly a way to study a factor when you want to adjust for some kind of relationship to another factor. One conclusion from a logistic regression analysis is what effect the variable of interest has on the outcome in the population when all other variables included are held fixed. That meant that I could use the logistic regression in a different way than I initially thought. Instead of adding one gene to the other in a nested way, I studied one gene at a time at first. What we tried to use the nested logistic regression for in the first paper was to adjust for the long range LD that is present in HLA. However, I turned it around a bit and thought that actually if there was long range LD between the genes that were of possible importance I should be able to pick those signals up within each one of the genes. After reducing the complexity within each gene, I could adjust for the other genes to remove hits that were due to LD with other more strongly associated alleles. The long range LD hits would then be redundant and the best signal should be kept by the model and those associated allele groups that were not part of long range LD would be adjusted for all other effects within our material and either hold for the adjustment or be excluded. In this way I reduced the number of variables in the final model, corrected for long range LD and at the same time the analysis was as unbiased as I could manage. However, the effects of *DRB1*15* and *A*02* were at this time more or less established and this complicated the concept a bit.

Another problem was under which model the allele groups should be coded in the analysis described above. We could no longer ignore the prior knowledge about *DRB1*15* and *A*02*. In our cohort, the OR for MS for those carrying one copy of *DRB1*15* is around 3, for the homozygotes the OR is much higher, around 11 in our cohort. This huge effect could not be found for *A*02*, which seemed to act in a more dominant fashion; the presence of any *A*02* allele reduced the risk of MS similarly as the presence of two *A*02* alleles. How all the other alleles should be coded was more or less impossible to predict. Therefore, I decided two things. First, I wanted a baseline model where I had the two already established risk factors for MS coded as correctly as possible given the previous results. Second, I wanted to study what the difference was between coding the rest of the alleles for carriage or dosage. Did it actually matter in my analysis? If there was no difference the model with fewer categories would be preferable due to reduced complexity.

I performed the analysis first with all allele groups coded as the number of alleles from a particular allele group (0, 1 or 2) and second with all allele groups coded as presence of any allele from the allele group (0 or 1). They were both compared with the same baseline and that baseline contained gender, country of origin, *DRB1*15* coded as number of *DRB1*15* alleles and *A*02* coded as presence or absence of *A*02* alleles. Both underlying hypotheses (carriage or dosage effect) included four sub analyses and one last final model. The first four tests were performed within each gene, i.e. one per gene. All allele groups within *DRB1* with a certain frequency was added to the baseline variables and those not significantly associated to MS was removed until only associated allele groups were left. When this was done for all four genes, all the alleles associated to MS were added into the final model to remove possible LD effects or at least adjust for it. Also here the non-significant allele groups were removed. I should perhaps mention that a sacrifice was made within each gene. I had to exclude one allele in each locus from the analysis, otherwise the first entered allele group within each gene was used as intercept whether I wanted it or not. My dataset did not contain any individuals that were “zeros”, i.e. didn’t have any allele group. Therefore I had to decide which allele group I wanted as intercept within each gene. After careful consideration I chose the allele group that was most equally distributed between cases and controls and had an allele frequency over some threshold that was fairly common in the cohort. This is something that can be debated, especially within *DRB1* since almost every major allele group at that gene has been reported to affect MS susceptibility in some way, as reviewed in [53].

4.1.6.3 Logistic regression revolution

When I had my two final models built on different underlying assumptions, the allele groups either acted dominantly or in a dosage fashion, I compared them to each other and discovered that both models contained the exact same allele groups. The estimates and the confidence intervals only had small differences and the overall fit of the model was a bit better with the carrier coded model. So the final result was the same and it didn’t seem to matter whether I coded alleles as carriage or dose. The main difference between the two assumptions was that dosage was a bit stricter in defining associated alleles within each gene while the carrier model allowed for more allele groups to be

associated with MS but these were later pruned from the final model due to LD with effects from other allele groups (Table 2).

Table 2. Comparison between the two different ways of coding the allele groups in the logistic regression.

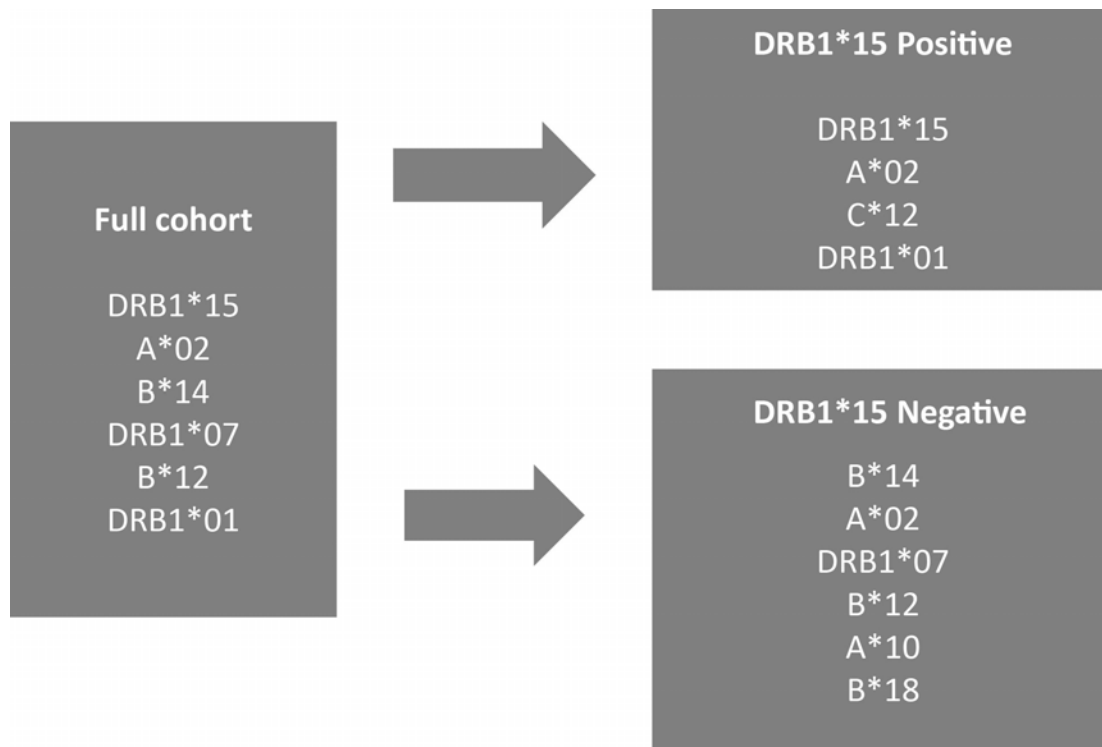
Dosage Model	Model terms, significant alleles	Model comparison	Δ Deviance	Δ Df	p-value
1	Baseline				
2	BL + B alleles (B*12, 14)	Model 2 vs.1	39.1	11	5.14x10 ⁻⁰⁵
3	BL + DRB1 alleles (DRB1*01, 07)	Model 3 vs.1	16	7	0.025
4	BL + A alleles (nas)	Model 4 vs. 1	5.4	6	0.5
5	BL + C alleles (nas)	Model 5 vs. 1	18.3	10	0.049
6	BL + B*12, B*14, DRB1*01, DRB1*07	Model 6 vs. 1	37.6	4	1.39x10 ⁻⁰⁷
Carriage Model	Model terms, significant alleles	Model comparison	Δ Deviance	Δ Df	p-value
1	Baseline				
2	BL + B alleles (B*07, 08, 12, 14, 17, 40)	Model 2 vs.1	44.9	11	5.07x10 ⁻⁰⁶
3	BL + DRB1 alleles (DRB1*01, 03, 05, 06, 07)	Model 3 vs.1	24.8	7	0.00083
4	BL + A alleles (nas)	Model 4 vs. 1	6.8	6	0.3
5	BL + C alleles (C*05)	Model 5 vs. 1	22.8	10	0.011
6	BL + B*12, B*14, DRB1*01, DRB1*07	Model 6 vs. 1	41.4	4	2.22x10 ⁻⁰⁸

Δ Df = delta degrees of freedom, indicates how many variables (allele groups) that differ between the two models compared. BL = baseline, nas = no allele group significantly associated to MS

I chose to use the data from the carriage model as primary results in paper II. In that final model we found association to MS of DRB1*01 and DRB1*07 within DRB1 and B*12 and B*14 on top of our baseline containing A*02 and DRB1*15. B*12 contains the subtypes B*44 and B*45. B*14 is in tight LD with C*08 which I found associated to MS in the first paper. B*14/C*08 is a quite rare allele group and I am not confident in the ability for others to replicate that finding. This association may very well be sample specific. The cohort in paper II is not suitable for replication of the findings in paper I since many individuals were included in both studies even though new subjects were added in the second paper. It is possible to confirm that A*02 is still associated with MS even after adjusting for HLA-B allele groups in this cohort.

Even though logistic regression adjust for DRB1*15 in this setting I wanted to see if any allele group had an independent effect within the DRB1*15 positive or negative groups. Therefor I stratified the cohort into two groups according to DRB1*15 status and performed a logistic regression within each group when the effect of carrying DRB1*15 was completely removed. In the DRB1*15 positive group DRB1*15 was still associated since several of the subjects were homozygous and as we already knew, this increases the risk of MS. That was not surprising, and the association to A*02 in each of the DRB1*15 strata was not surprising either but rather a nice confirmation of the independency of A*02 from DRB1*15. What was surprising however, was that the rest

of the allele groups that was found associated in the total cohort showed association to MS in one of the strata, never both (Figure 3).



*Figure 3. Comparison of how the allele groups associated to MS in the full cohort were associated within the DRB1*15 positive and negative strata. The right side squares also display which allele groups that reached significant association to MS in this setting but was not found in the full cohort. Figure based on Table 1 and Table 2 in paper II where details of odds ratios, p-values and frequencies can be found.*

This result prompted us to investigate possible interactions with *DRB1*15*. *DRB1*01* was seen to decrease the risk of *DRB1*15* more than was expected from the overall effect of *DRB1*01*. The same trend was seen for *C*12* and somewhat also for *A*02*. It is hard to say exactly what this type of interaction means. Obviously the carriage of, for example, both *DRB1*01* and *DRB1*15* does not resemble the expected risk modification within this cohort but if this result is due to an actual physical interaction resulting in attenuation of risk to MS would be nothing but speculation. The interaction only tells me that these two factors do not follow the expected rules I set up in the model. Subjects might have more use of this combination of class II allele groups when fighting some pathogen or complement each other with other factors also present on these haplotypes. It doesn't have to be *DRB1*01* at all, but rather something in LD with it. *DRB1*01* is perhaps more expressed or even lowers the expression of *DRB1*15* thereby reducing its influence in MS susceptibility.

When repeating the analysis found in the Italian paper, adjusting for allele groups and studying one variable at the time [52], I could not replicate the finding where *C*05* enhanced the protection of *A*02*. The overall result from that analysis in this cohort was in fact that *C*05* seemed to have no effect what so ever when adjusting for *B*12* or *A*02* or both.

4.1.7 Haplotypes of allele groups

When trying to estimate the haplotypes in the HLA region I encountered annoying problems. At first the program UNPHASED 3.1.4 had the maximum limit of only 16 alleles. Normally, this is no problem, but for me it was crucial to use all the information I had on each individual. That meant that I did not group the less frequent allele groups in each gene into one “X”-group, and when that was entered in UNPHASED 3.1.4 it refused my kind commands. I had to use an earlier version and the most recent one that accepted a higher number of alleles was UNPHASED 3.0.13. Then the next problem appeared, my computer was not equipped to handle the amount of RAM the process consumed. A colleague of mine had a computer more suitable for the analysis but still the software complained, the data was too complex for the standard settings which had to be adjusted; the software then did the calculations in approximately a week.

The reliability of each estimated haplotype is given by the estimation procedure and I only used those individuals whose probability were greater than 80%. The results were nonetheless almost the same when including all haplotypes or only those with high probability. Individuals that were excluded mostly had unusual haplotypes of which I wouldn't be able to draw any conclusions anyway. In my first analysis I wanted to see if any of the 20 most common haplotypes in the cohort affected MS susceptibility. Logistic regression was used to adjust for carriage of more than one of the most common haplotypes. Indeed most of the associated haplotypes contained one or more of the alleles that were seen associated on their own. What could easily be seen in Table 3, was that the haplotypes carrying *DRB1*15* was more common in patients than in controls and most exhibit significant association to MS. Also *A*02* showed a pattern, being mostly protective, but only as long as the haplotype did not carry *DRB1*15*. Somehow *A*02* do not decrease the effect of *DRB1*15* on its own but in concert with *C*05* and *B*44* which is a neutral haplotype in our cohort. The question was if these three allele groups were adding protection by simply being present in one individual, if it was the haplotype in itself or if they tagged something else that was important in MS susceptibility.

*Table 3. (Adapted from Paper II: Table 4). The table shows the most common haplotypes in the cohort, all entered into a logistic regression to account for an individual carrying two of them. Neutral DRB1*15 haplotype is underlined.*

No.	Haplotype				Frequency Cases (%)	Frequency Controls (%)	FDR corrected p-value	Odds Ratio (95% CI)
	HLA-A	HLA-C	HLA-B	DRB1				
1.	1	7	8	3	7.0	8.6	0.033	0.78 (0.63-0.96)
2.	3	7	7	15	7.6	4.2	2.28x10 ⁻⁰⁶	1.90 (1.49-2.44)
3.	2	7	7	15	6.0	2.1	2.72x10 ⁻¹¹	3.14 (2.31-4.34)
4.	2	3	15	4	1.8	3.1	0.015	0.60 (0.42-0.87)
5.	3	4	35	1	1.4	2.6	0.0087	0.54 (0.36-0.81)
6.	2	5	12	4	0.8	2.9	1.60x10 ⁻⁰⁶	0.28 (0.17-0.45)
7.	9	7	7	15	2.6	0.8	8.33x10 ⁻⁰⁶	3.30 (2.06-5.47)
8.	2	3	40	6	1.3	2.0	0.097	0.67 (0.43-1.03)
9.	10	12	18	15	1.5	0.8	0.021	2.00 (1.19-3.48)
10.	3	3	15	4	1.1	0.5	0.016	2.35 (1.27-4.52)
11.	1	6	37	15	1.2	0.4	0.00095	3.87 (1.97-8.35)
12.	2	5	12	15	0.8	0.8	0.82	0.92 (0.51-1.69)
13.	1	7	7	15	1.0	0.5	0.032	2.15 (1.15-4.19)
14.	11	4	35	1	0.7	0.8	0.68	1.19 (0.63-2.23)
15.	19	16	12	7	0.7	0.8	0.82	0.93 (0.49-1.74)
16.	19	3	40	4	0.4	1.1	0.015	0.38 (0.18-0.74)
17.	2	7	8	3	0.5	0.9	0.10	0.56 (0.29-1.05)
18.	2	3	40	4	0.6	0.9	0.26	0.65 (0.34-1.25)
19.	2	3	15	6	0.5	0.9	0.28	0.67 (0.34-1.27)
20.	3	7	7	4	0.5	0.6	0.72	0.85 (0.42-1.71)

FDR = false discovery rate, correction of the p-value to account for the probability of false positive findings CI = confidence interval, an estimation of in which range the populations' OR would be found.

I tried to answer these questions by first studying the independence of the haplotype to *DRB1*15* and found that it was independent. I investigated the importance of the class I allele groups presence on protective ability of the haplotypes. *A*02* together with *B*44* was enough to reach the same level of protection as the full haplotype although the OR was a bit lower when also *C*05* was present on the chromosome. Finally, the change in OR between carrying only the haplotype, *DRB1*15* and both was studied. *DRB1*15* carrying haplotypes had an increased risk of disease, haplotypes carrying *A*02*, *B*44* and *C*05* had a decreased risk of disease and haplotypes positive for both *DRB1*15*, *A*02*, *B*44* and *C*05* did not differ in risk of disease from those not carrying *DRB1*15* or the full protective haplotype.

The conclusion, that the haplotype of *A*02*, *C*05* and *B*44* is protective in MS, needs to be verified in some other cohort. To my knowledge, this approach has only been performed once before in a case control material in MS, and that was in Sardinia in 2001 [54]. That population is genetically more isolated and the association pattern of HLA to MS is different than in northern Europe and my study cannot really be used as a replication cohort in that sense. It would be nice to investigate this haplotype in some cohort more similar in genetic etiology.

4.1.8 Perspectives from papers I and II

I have already in the text above discussed some issues that had to be dealt with when doing these two studies, especially regarding the most appropriate way to analyze the association to HLA in a complex disease. Given the result from my two studies and the information others have published I would say that two things are more important than others when going further with these studies. First, my associations are based on allele groups. Within a group of haplotypes, the protection might be located only within a subgroup of these alleles. When I create haplotypes with allele groups on other genes, I study a subgroup within the allele group meaning that it might be a sub allele group within either *A*02* or *B*44* that harbor the protection to MS that just happens to be refined when adding info on *HLA-C* for example. Therefore the analysis of alleles within this cohort is warranted. However, my second thought is that it could also be possible that the protection to MS is not located within the classical HLA class I genes at all but something else on the chromosomes positive for *A*02* and *B*44*. This would explain why the LD to different allele groups in other populations is following similar but still show a bit different patterns, in Italy more linked to *A*02* and the *A*02-C*05* haplotype and in Britain more linked to *B*44* and/or *C*05*. In Canadian family studies no signals was seen in class I at all after correcting for *DRB1* allele groups [55]. However there is a debate within the field that the role of class I and class II may differ between populations due to different LD structures and this was highlighted in the commentary that followed the Healy paper where *B*44* was found to be protective in certain MRI outcomes and MS susceptibility [56].

I am not convinced that I succeeded in finding the answer to if there are one or two protective effects in class I. I would like to rule out the role of *C*05* as being protective in itself but the role of refining a protective haplotype within the allele groups *A*02* and *B*12* is still possible. My results can however not distinguish if *A*02* and *B*12* are totally independent protective signals or if they are part of the same signal, it seems to be a combination of both which is a bit confusing and frustrating for me. When correcting for *DRB1*15* and *A*02*, *B*12* is still significantly associated to MS, hence an independent signal. *C*05* is not associated to MS at all when correcting for either *A*02* or *B*12*, hence not independent. When studying haplotypes with both *A*02* and *B*12* the odds ratio is not that different from that of the *A*02-C*05-B*12* haplotype. But the haplotypes carrying only *A*02* or *B*12* both have significantly higher risk than the protective haplotype; hence the effect of having them on the same haplotype is more beneficial for MS risk.

4.2 PAPER III

In February 2009, a study from Ebers's group about a vitamin D response element in the promoter region of the *DRB1*15:01* allele was published and this was the basis to the collaboration and publication of paper three in this thesis [57]. The background to this story was that a month of birth effect had been seen in MS that suggested that individuals born in the spring had a higher risk of MS later in life and those born in late fall had an decreased risk of MS, this has later been observed in both hemispheres [58-61]. As the interest for the vitamin D hypothesis in MS literally exploded within our field, the paper from Ebers group received much attention. It was the perfect explanation for everything! The geographical gradient in MS risk, here explained by

degree on sunlight, interacting with the largest genetic risk factor (*DRB1*15*) was something that would put some pieces together in the MS puzzle. It was in this light we did this study. Looking back, reading all the criticism around the publication that was part of the basis for this study, I realize that paper III might have been based on something neither true nor replicable [32, 62, 63]. Today, I would hesitate to draw the same conclusions as then and I would have analyzed it in a different way.

The basic assumptions for this study were that *DRB1*15* bearing haplotypes harbor a vitamin D response element that regulates expression of the gene. The hypothesis was that vitamin D deficiency during pregnancy or early childhood could affect the risk of MS later in life. In the paper we saw an increase of April births among the *DRB1*15* positive patients compared to the *DRB1*15* negative patients. We also saw a decrease in births in November for *DRB1*15* positive patients. These effects could not be seen within the controls.

If repeating the study, I would look for MS risk within *DRB1*15* positive and MS risk within *DRB1*15* negative for each month, not the other way around as we did in the paper. In this way we also would correct for a possible effect of *DRB1*15* on month of birth in general, which might be a confounder. Today, my first analysis would be to check if there is a month of birth effect at all in this cohort. Second, I would assess the power of the study to detect a difference similar to the ones in the larger studies such as the one from Willer et al [61]. If we have sufficient power, an analysis of a possible *DRB1*15* effect could be warranted. If we don't have power even when not splitting the material up, I would suggest a careful interpretation of the results and correction of p-values. This was not done in the published paper.

After this paper was published, there have been several studies about vitamin D and factors surrounding that pathway in MS. The hypothesis that vitamin D levels influence expression of *DRB1*15* and thereby the risk of MS have shifted towards the idea that lower levels of sunlight can increase risk of MS independent of HLA alleles, as was outlined in an Australian meta-analysis [64]. Also, a month of birth effect was not found in a large GWAS [32] and a month of birth effect was not connected to *DRB1*15* in Finland [62]. The degree of reduced sunlight exposure further away from the equator coincide with an increased population frequency of the *DRB1*15* allele and MS prevalence, thereby making it hard to distinguish which is the primary factor responsible for this effect. An investigation of the relationship between vitamin D and carriage of *DRB1*15* did not see any interaction affecting on MS risk [63]. The authors did however use the vitamin D levels in adulthood at, or close to, time of diagnosis and these may differ to the levels in childhood.

4.3 PAPER IV

The idea behind paper four was presented to me by my colleagues Malin, Katharina and Anna who already had an idea around a research question that addressed NAb (neutralizing antibody) development in IFN- β (interferon beta) treated patients. A group in Germany had published that NAb development in MS patients treated with IFN- β differed with *HLA-DRB1* genotype [65-67], and the idea was to study this also

within our cohort since we have slightly different outcome measures. The design was however not that easy, which we realized quite soon.

The German studies used a different technique to study NAb presence in serum from MS patients on IFN- β treatment than what is used in the Swedish NAb registry. Their technique detects not only NABs but also BAb (binding antibodies) directed to various epitopes on the IFN- β molecule that does not necessarily abolish the binding ability of IFN- β to its receptor. Thereby, the BAb do not always interfere with the biological function of the protein which is measured by the downstream expression of MxA in real time PCR. Their three outcomes were therefore 1) presence of antibodies targeting IFN- β , 2) presence of neutralizing ability among those antibodies and 3) the percentage of neutralizing ability of the antibodies (Table 4).

With our technique, the first outcome was NAb positivity since we measured MxA expression as a first step in the NAb analysis. The next outcome was the presence of clinically relevant titers of NABs defined as the titer necessary to abolish the biological activity of IFN- β in vivo. Our third outcome was the actual titer level and to see if any allele group affected the ability of producing high titers. The outcomes of our and the German studies are illustrated in figure 4 and summarized in table 4.

Table 4. Definitions of different outcomes used in the German studies [65-67] and our study.

Outcome German studies	Definition
All AB	Binding ability of antibody to IFN coated ELISA
Biologically active antibodies (NABs)	Enough Abs to reduce MxA expression >50%
Neutralizing antibody activity	Point estimation of neutralizing ability of antibodies in each sample
Outcome our study	Definition
Nab	Presence of NABs (>10 TRU/ml)
Clinically relevant titers	Enough NABs to reduce MxA expression ~ 80% (>150 TRU/ml)*
Titer level of Nabs	Point estimation of NAb titer in each sample

*TRU = Ten fold reduction units, IFN = interferon, AB = antibody, * = as reported in [68].*

The noise of irrelevant binding of BAb was naturally reduced by the nature of the MxA. The choice of analyzing a previously reported relevant cut off for neutralizing ability [68], increases the clinical importance of this study.

German studies

Our study

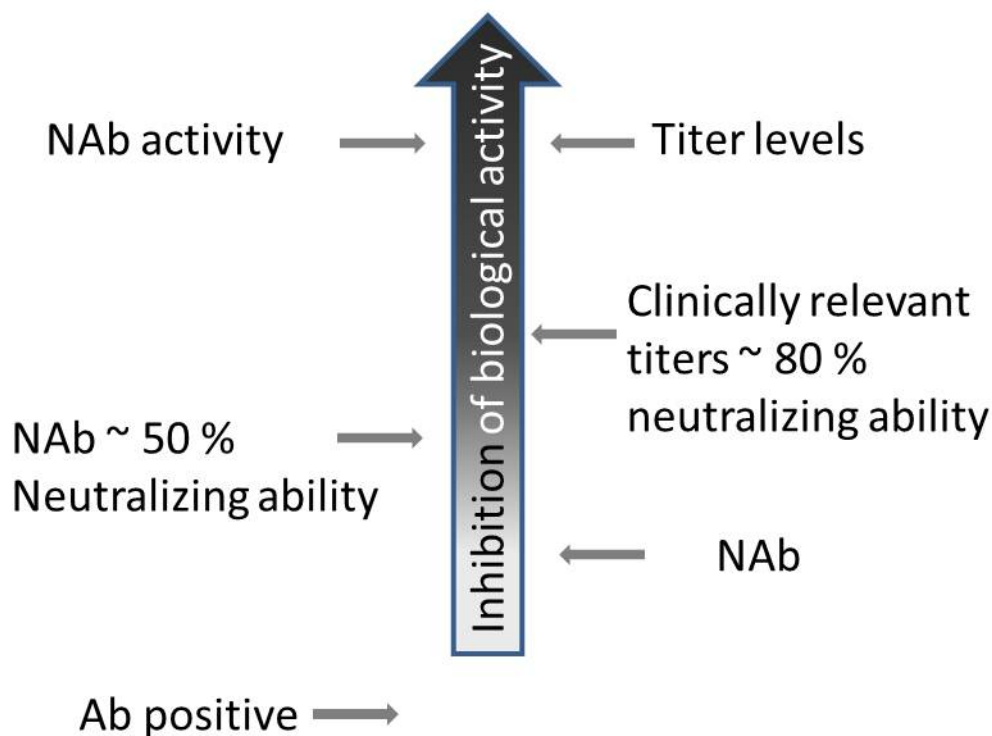


Figure 4. Comparison of level of inhibition of biological activities in the outcomes. NAb activity spans everything from 0 to 100% neutralizing ability in the German studies, titer level in our study spans from 10 TRU/ml to 65,000 TRU/ml.

After a lot of discussions, it was decided that we should analyze all patients that were positive for NABs and the patients that was truly negative. We thereby ignored those that had no samples taken within 36 months from beginning of treatment. However, not all the patients fulfilling the criteria within the NAb registry were HLA genotyped. Since the main question was to determine the effect of HLA genes in development of NABs we reduced the sample population to those patients that were HLA typed, mainly patients included in paper I and II in this thesis.

4.3.1 Patient ascertainment biases

There are three main biases in the patient ascertainment for this study and this type of problem is hard to avoid in clinical research.

- 1) Patients who more often visit the clinic have a higher probability of being asked to donate a sample of blood for research (and thereby being HLA-genotyped). They also have a larger set of data points in our registries which increase their probability to fulfill the criteria of this study. Why do patients go to the clinic often? They might have more symptoms, either due to a more severe disease course or due to no effect of the treatment. This inflates the number of patients in the NAb registry with a higher probability of NABs.
- 2) The Swedish NAb registry was started in 2003 and NAb tests are taken around 12 and 24 months after start of treatment. Other samples are taken at other time points either as follow up after a positive test, or if there seems to be no effect

of the treatment. The patients we defined as negative in the study have all been tested at the clinic from 2003 and onwards. Those who have a positive sample might have been positive for a long time previous to the test, while older patients negative for a long time were excluded due to absence of data points within the first 36 months of treatment. Therefore the NAb positive patients are older than the NAb negative patients in this study and this could influence the ability of producing NABs.

- 3) Between the years 1993 and 1998, Betaferon was the only IFN- β treatment used in the clinic. Later Avonex and Rebif was introduced, but for older patients closer to conversion to SPMS, the high dose Betaferon was used to a higher degree since it had better effect on relapse inhibition despite more side effects. Avonex or Rebif are given to younger patients with a slower progression, since these drugs are administered fewer times per week and therefore affect daily life less. Many patients therefore started with Betaferon, became NAb positive and experienced no effect of the treatment. When Avonex and Rebif were introduced the patients who switched could remain positive for NABs due to cross reactivity from previous Betaferon use [69-71]. This would give us a bias of higher probability of positivity connected to older patients who used Betaferon when Avonex and Rebif were not used. These patients may be registered on another treatment today.

These biases call for somewhat cautious evaluation of the results and a role of potential confounding is warranted. The effect of treatment regimen on NAb development has previously been studied within the full registry data, showing that Betaferon and Rebif induce NAb titers to a larger extent than Avonex [72] which was also seen in our sample. However, given the conservative inclusion criteria for the negative group in this design, the effect on NAb induction is inflated.

4.3.2 Genetic impact on NAB development

In this study, we could see a weak association between *DRB1*15* and NAb development which became stronger if correcting for treatment. An increased OR for NAb development in Rebif users if positive for *DRB1*15* compared to those that were *DRB1*15* negative was found. This was also shown for titers high enough to diminish the effect of the treatment. It would only be logical if an association of a HLA allele to antibody development were of class II origin since CD4⁺ T-cells which recognize peptides on HLA class II molecules are excellent activators of antibody producing B-cells. However, in Betaferon treated patients we saw an increase in odds of NABs if the patients carried the class I allele group *HLA-A*02*. This result must be verified in another cohort because it is based on a test of 42 *A*02* positive NAb positive patients and only one *A*02* positive NAb negative patient that could result from either a clinically relevant effect (however difficult to explain) or be a sample bias.

The development of NABs was also higher if treated with Rebif or Betaferon while Avonex does not induce NABs to the same extent. Would it be wise to put all patients on Avonex to start with, and extra beneficial for those who are *DRB1*15* positive? This

is a very strong conclusion based on a study that is to a certain extent biased and not representative for the whole clinic base of patients. That is important to remember.

The German studies identified an association between *DRB1*04* and NABs, and also subtyped the *DRB1*04* allele group to identify if all alleles were associated with risk. *DRB1*04:01* and *DRB1*04:08* was associated with higher risk while *DRB1*04:04* was not. In our study we did not see an association between *DRB1*04* and NAb development and this can be the result of higher frequency of *DRB1*04:04* in our population. This is something that I would like to study further in this paper.

5 FUTURE PERSPECTIVES

My results from the four studies do add some information to the previously known data on HLA in MS susceptibility. We now have further evidence that protection to MS does not only involve *HLA-A*, but also *HLA-B* probably in combination. We know that there do seem to be a month of birth effect in MS even if we could only see it when stratifying on *DRB1*15*, however the link to vitamin D is a bit vague. And lastly there seems to be a modulating effect of *DRB1*15* on development of NABs but it does not overcome the effect of the treatment regimen themselves.

Since we have some previous clues to what actual role HLA have in MS the hypotheses of possible interactions are numerous and need to be studied in detail. Both interactions between *DRB1* alleles, in-between genes and with environment are important since they are all involved in the main hypotheses of MS immunopathogenesis. The immune system is built to interact with and respond to stimuli in the surrounding milieu. Disease susceptibility might depend on if HLA molecules can bind certain peptides more efficiently like in celiac disease [19] or if certain HLA alleles have differential expression, thereby presenting different amount of antigens. That would just add another level of complexity to the question. DNA methylation of *DRB1*15* does not seem to affect MS severity [73] but that does not rule out that other haplotypes carry important methylation patterns. Perhaps whole blood is not the most appropriate tissue to study for functional studies in MS as the methylation pattern for a gene can vary from tissue to tissue [74]. CSF derived cells are naturally a more suitable source for this type of study, but the amount is limited. It would also be valuable to study one cell type at the time since the relative amount of immune cells in blood is different between individuals.

The role of genetics on treatment effect is quite interesting since it would enable clinicians to evaluate the effectiveness of different treatments in a patient. The other way around, that type of information could be used to develop better treatments since it is a clue to how the body reacts to the drug. This is however only true if all other parameters are held fixed, having the same side effects, same distribution pattern and so forth. As each individual is unique there are always those that will react differently to the same drug even if we can control for a specific HLA allele group. When considering the impact of the genetic effect compared to the drug effect in this study I am not convinced that this will be applicable in the clinic, however it might point in which direction to go when trying to eliminate the risk of having no effect of the treatment.

I think the next step to study the role of HLA in MS is to learn from other autoimmune diseases and go back to basic studies of expression, binding and interaction with the T-cell receptors especially in settings of environmental factors. And perhaps not only study *DRB1*15*, who knows what we will find if we broaden our mind a bit and study also the protective alleles like *DRB1*01* or *DRB1*07* in these settings or the other polymorphic genes included in the risk haplotype.

One way of getting closer to the role of HLA in MS could be to study if certain amino acids or motifs of the HLA molecules are more strongly associated to disease than the classic *DRB1*15* haplotype. That would indicate an importance of the HLA proteins in MS susceptibility.

6 ACKNOWLEDGEMENTS

A lot of people have helped me through the time that led to this thesis; there are some individuals that I would like to thank for that.

Jan Hillert for being a terrific supervisor by letting me take the time I needed to find my path into the HLA complexity. Without all the detours and “eureka”-moments I wouldn’t have learned so much. The support and interest in my numerous analyzes, results and the patience with my confusion about what the results really meant has been greatly appreciated :)

Ingrid Kockum, we have a history that goes beyond this thesis. Thank you for believing in me when you recommended me for this doctoral position. Your knowledge of genetics in complex disease and your ability to make me understand it has been priceless.

Boel Brynedal, once the “maker”, later the supervisor :) When I first came to the group you tried to teach me the complexity of HLA but I just didn’t get it. We had numerous conversations about haplotypes of alleles which also actually was haplotypes of other alleles. My first thoughts were that the world was upside down in Huddinge compared to Solna and I felt very confused... But soon I started to get the hang of it and you guided me through these years with confidence that I could do this my way.

“The Huddinge Gang”

Boel Brynedal, Iza Lima, Kerstin Imrell, Frida Lundmark and Tomas Masterman. I would like to thank you all as a group for making the time in Huddinge so fantastic! We had so many interesting and inspiring conversations and our whiteboard was always full of drawings, tables and abbreviations that nobody other than us understood. The fika-discussions could be both joyful and challenging when trying to understand statistics, epidemiology, pedigrees, interaction, correction for multiple testing and linkage disequilibrium. Everybody had their own opinion! Thank you for letting me into that atmosphere of knowledge!

Kerstin Imrell, the way you think of science is so different from my way of thinking. You have forced me to think about other aspects of my projects and it makes you a terrific source of inspiration. Thank you for all the conversations about life and relationships and for making me see things from a different perspective.

Iza Lima, is there anything we have forgot to discuss? I think not. I have loved all our inspiring talks about all or nothing, both gossiping and serious business. Our endless tries to finally deal with this multiple comparison thingy that just refuses to be logical at any point! We have skied together in Santa Fe and hitchhiked together with Kerstin in Turin. Thank you for all the help in making me understand statistics and making me see that you can learn these things if you go deep enough and really dig into it. Apparently nothing is impossible as long as there is still an unread textbook in the subject. Without you, this thesis would have been so much thinner!

I would also like to thank the rest of the MS genetics group. **Wangko Lundström**, for your interest in science and your sense of humor. **Helga Westerlind**, for your endless knowledge in which computers and mobile telephones not to buy, and for your support when I needed more RAM for my data analysis! **Katharina Fink**, **Eva Greiner** and **Virginija Karrenbauer** for your knowledge in neurology and making me understand the clinic and patient data much better. **Anna Glaser** and **Ryan Ramanujam**, the newcomers in the group, for all the fresh ideas you bring with you.

The immunology group, with a special thanks to **Anna Fogdell Hahn**, **Malin Lundkvist** and **Christina Hermanrud** for the collaboration and support with the fourth manuscript in this thesis and for reading my thesis. Thank you **Malin** for the skiing companionship in Stöten, we had a very nice time until I crashed into that tree with my snowboard :) Thank you **Rasmus Gustafsson**, **Elin Engdahl**, **Roger Jungedal**, **Ajith Sominanda** and **Clemens Warnke** for all the inspiring chats and help in the lab. **Jenny Ahlqvist** for starting the tradition of the wandering turquoise vase and for giving me excellent input on my thesis.

Marjan Jahanpanah, for all the lovely food and joyful conversations in Huddinge. I miss your energy and warmth in Solna! **Annelie Porsborn** for making me want to discover new TV-series and for making the lunches in Huddinge so nice! **Merja Kanerva**, **Anna Mattsson** and **Ingegerd Löfving-Arvholm** for much appreciated help in the lab and making the lab work so much more fun. **Cecilia Svaren-Quiding** for all our endless discussions about both this and that, solving the mysteries of HLA genotyping and trying to make everybody keep the lab nice and tidy! **Leszek Stawiarz** and **Karin Lycke** for your endless knowledge about patient registries.

Emilie Sundqvist for your interesting sense of humor and for our computer game discussions. Your need for information is overwhelming ☺ **Magdalena Lindén**, for your passion for science and your endless help with BC gene data files, **Magnus Lekman** for teaching me how to use Illustrator when I needed it the most and for always asking me if I am going to the gym later when you catch me eating “mellanmål” in the afternoon ☺ **Cecilia Dominguez** for always smiling and wanting to know more about HLA. **Samina Asad** and **Alexandra Gyllenberg**, we met many years ago and had quite a lot of fun on floor 00 before I went to Huddinge. Good luck with your theses.

CMM colleagues from floor 00, 02, 04 and 05 for interesting discussions during the years, with a special thanks to **Tomas Olsson**, **Maja Jagodic**, **Johan Öckinger** and **Melanie Thessen-Hedreul** (good luck with your thesis). **Amennai Beyeen**, thank you for all the help with “selecttjänsten”. **Magnus Nordenskjöld**, **Elisabeth Ekelund** and **Maria Bradley**, for providing the scientific environment that started my scientific journey. **Selim Sengül**, for teaching me how to use the techniques in the lab when I first came to CMM. **Rudolf**, **Dagmar**, **Daniel** and **Olle** for all the help with IT problems.

I would like to thank all the **co-authors of the papers**, and especially those **from Oslo** for your support in paper I and II and the **Nordic MS genetics group** for providing a nourishing environment for research.

Tack till alla härliga **korvdansare** som samlas på Sweden Rock Festival varje år! Det är ett härligt avbrott i vardagen att umgås med er.

Lisa Israelsson, utan dig hade det inte blivit någon avhandling. Tack för att du livade upp åren på Stockholms Universitet med din galna humor. Din passion för immunologi och att alltid vilja veta varför något sker har hjälpt mig mycket. **Victoria Olausson**, för din otroliga öppenhet och vilja att bjuda in mig i ditt liv när vi bodde i korridoren tillsammans. Du har fått mig att fundera över mycket i livet.

Johanna Björling, tack för att du stöttar mig i vardagen och hindrar mig från att ta för snabba beslut. Vem vet vad jag hade gjort idag om inte du hade funnits.

Mamma, för din förmåga att ta ner mig på jorden när jag svävar för högt. **Johan**, för allt vi gått igenom i livet och din underbara humor som alltid är ett vattenhål i vardagen. **Peter**, för att du förstår mig mer än någon annan.

7 REFERENCES

1. *Finishing the euchromatic sequence of the human genome*. Nature, 2004. **431**(7011): p. 931-45.
2. Lander, E.S., et al., *Initial sequencing and analysis of the human genome*. Nature, 2001. **409**(6822): p. 860-921.
3. Stoneking, M. and J. Krause, *Learning about human population history from ancient and modern genomes*. Nature reviews. Genetics, 2011. **12**(9): p. 603-14.
4. Gabriel, S.B., et al., *The structure of haplotype blocks in the human genome*. Science, 2002. **296**(5576): p. 2225-9.
5. Dudbridge, F., *Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data*. Hum Hered, 2008. **66**(2): p. 87-98.
6. Haines, J.L.P.-V., M. A., *Genetic analysis of complex disease* 2006, Hoboken, New Jersey: John Wiley & Sons, Inc.
7. Rothman, K., *Epidemiology: an introduction* 2002, New York: Oxford University Press.
8. Ahlgren, C., A. Oden, and J. Lycke, *High nationwide prevalence of multiple sclerosis in Sweden*. Multiple sclerosis, 2011. **17**(8): p. 901-8.
9. Neefjes, J.J. and H.L. Ploegh, *Allele and locus-specific differences in cell surface expression and the association of HLA class I heavy chain with beta 2-microglobulin: differential effects of inhibition of glycosylation on class I subunit association*. European journal of immunology, 1988. **18**(5): p. 801-10.
10. Neefjes, J., et al., *Towards a systems understanding of MHC class I and MHC class II antigen presentation*. Nature reviews. Immunology, 2011. **11**(12): p. 823-36.
11. Vigneron, N., et al., *An antigenic peptide produced by peptide splicing in the proteasome*. Science, 2004. **304**(5670): p. 587-90.
12. Neisig, A., et al., *Allele-specific differences in the interaction of MHC class I molecules with transporters associated with antigen processing*. Journal of immunology, 1996. **156**(9): p. 3196-206.
13. Nimmerjahn, F., et al., *Major histocompatibility complex class II-restricted presentation of a cytosolic antigen by autophagy*. European journal of immunology, 2003. **33**(5): p. 1250-9.
14. Plantinga, T.S., et al., *Modulation of inflammation by autophagy: consequences for Crohn's disease*. Current opinion in pharmacology, 2012.
15. Olerup, O. and H. Zetterquist, *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, 1992. **39**(5): p. 225-35.
16. Steck, A.K. and M.J. Rewers, *Genetics of type 1 diabetes*. Clinical chemistry, 2011. **57**(2): p. 176-85.
17. Bax, M., et al., *Genetics of rheumatoid arthritis: what have we learned?* Immunogenetics, 2011. **63**(8): p. 459-66.
18. Deng, Y. and B.P. Tsao, *Genetic susceptibility to systemic lupus erythematosus in the genomic era*. Nature reviews. Rheumatology, 2010. **6**(12): p. 683-92.
19. Sollid, L.M. and B. Jabri, *Celiac disease and transglutaminase 2: a model for posttranslational modification of antigens and HLA association in the*

- pathogenesis of autoimmune disorders*. Current opinion in immunology, 2011. **23**(6): p. 732-8.
20. McDonald, W.I., et al., *Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis*. Ann Neurol, 2001. **50**(1): p. 121-7.
 21. Poser, C.M., et al., *New diagnostic criteria for multiple sclerosis: guidelines for research protocols*. Ann Neurol, 1983. **13**(3): p. 227-31.
 22. Compston A., M.W.I., Noseworthy J., Lassmann H., Miller D., Smith K.J., Wekerle H., Confavreux C., *McAlpine's Multiple Sclerosis 4th Edition* 2005, London: Elsevier Inc.
 23. *PRISMS-4: Long-term efficacy of interferon-beta-1a in relapsing MS*. Neurology, 2001. **56**(12): p. 1628-36.
 24. Jacobs, L.D., et al., *Intramuscular interferon beta-1a for disease progression in relapsing multiple sclerosis. The Multiple Sclerosis Collaborative Research Group (MSCRG)*. Annals of neurology, 1996. **39**(3): p. 285-94.
 25. T.I.M.S.S., *Interferon beta-1b is effective in relapsing-remitting multiple sclerosis. I. Clinical results of a multicenter, randomized, double-blind, placebo-controlled trial. The IFNB Multiple Sclerosis Study Group*. Neurology, 1993. **43**(4): p. 655-61.
 26. Hillert, J. and O. Olerup, *Multiple sclerosis is associated with genes within or close to the HLA-DR-DQ subregion on a normal DR15,DQ6,Dw2 haplotype*. Neurology, 1993. **43**(1): p. 163-8.
 27. Jersild, C., *Studies of HLA antigens in multiple sclerosis*. Bollettino dell'Istituto sieroterapico milanese, 1978. **56**(6): p. 516-30.
 28. Jersild, C. and T. Fog, *Histocompatibility (HL-A) antigens associated with multiple sclerosis*. Acta Neurol Scand Suppl, 1972. **51**: p. 377.
 29. Brynedal, B., et al., *HLA-A confers an HLA-DRB1 independent influence on the risk of multiple sclerosis*. PloS one, 2007. **2**(7): p. e664.
 30. Fogdell-Hahn, A., et al., *Multiple sclerosis: a modifying influence of HLA class I genes in an HLA class II associated autoimmune disease*. Tissue antigens, 2000. **55**(2): p. 140-8.
 31. Harbo, H.F., et al., *Genes in the HLA class I region may contribute to the HLA class II-associated genetic susceptibility to multiple sclerosis*. Tissue Antigens, 2004. **63**(3): p. 237-47.
 32. Sawcer, S., et al., *Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis*. Nature, 2011. **476**(7359): p. 214-9.
 33. Lassmann, H., *Multiple sclerosis: is there neurodegeneration independent from inflammation?* Journal of the neurological sciences, 2007. **259**(1-2): p. 3-6.
 34. Stadelmann, C., C. Wegner, and W. Bruck, *Inflammation, demyelination, and degeneration - recent insights from MS pathology*. Biochimica et biophysica acta, 2011. **1812**(2): p. 275-82.
 35. Herz, J., F. Zipp, and V. Siffrin, *Neurodegeneration in autoimmune CNS inflammation*. Experimental neurology, 2010. **225**(1): p. 9-17.
 36. Chastain, E.M. and S.D. Miller, *Molecular mimicry as an inducing trigger for CNS autoimmune demyelinating disease*. Immunological reviews, 2012. **245**(1): p. 227-38.
 37. Sundqvist, E., et al., *Epstein-Barr virus and multiple sclerosis: interaction with HLA*. Genes and immunity, 2012. **13**(1): p. 14-20.
 38. Lucas, R.M., et al., *Epstein-Barr virus and multiple sclerosis*. Journal of neurology, neurosurgery, and psychiatry, 2011. **82**(10): p. 1142-8.

39. Raddassi, K., et al., *Increased frequencies of myelin oligodendrocyte glycoprotein/MHC class II-binding CD4 cells in patients with multiple sclerosis*. Journal of immunology, 2011. **187**(2): p. 1039-46.
40. Luckey, D., D. Bastakoty, and A.K. Mangalam, *Role of HLA class II genes in susceptibility and resistance to multiple sclerosis: studies using HLA transgenic mice*. Journal of autoimmunity, 2011. **37**(2): p. 122-8.
41. Ramagopalan, S.V., J.C. Knight, and G.C. Ebers, *Multiple sclerosis and the major histocompatibility complex*. Current opinion in neurology, 2009. **22**(3): p. 219-25.
42. IMMSGC, I.M.S.G.C., *A second major histocompatibility complex susceptibility locus for multiple sclerosis*. Ann Neurol, 2007. **61**(3): p. 228-36.
43. D'Alfonso, S., et al., *A sequence variation in the MOG gene is involved in multiple sclerosis susceptibility in Italy*. Genes Immun, 2008. **9**(1): p. 7-15.
44. Burfoot, R.K., et al., *SNP mapping and candidate gene sequencing in the class I region of the HLA complex: searching for multiple sclerosis susceptibility genes in Tasmanians*. Tissue Antigens, 2008. **71**(1): p. 42-50.
45. Cree, B.A., et al., *A major histocompatibility Class I locus contributes to multiple sclerosis susceptibility independently from HLA-DRB1*15:01*. PLoS ONE, 2010. **5**(6): p. e11296.
46. De Jager, P.L., et al., *Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci*. Nat Genet, 2009. **41**(7): p. 776-82.
47. Payami, H., et al., *Relative predispositional effects (RPEs) of marker alleles with disease: HLA-DR alleles and Graves disease*. American journal of human genetics, 1989. **45**(4): p. 541-6.
48. Dymont, D.A., et al., *Complex interactions among MHC haplotypes in multiple sclerosis: susceptibility and resistance*. Hum Mol Genet, 2005. **14**(14): p. 2019-26.
49. Ramagopalan, S.V., et al., *The inheritance of resistance alleles in multiple sclerosis*. PLoS Genet, 2007. **3**(9): p. 1607-13.
50. Nejentsev, S., et al., *Localization of type I diabetes susceptibility to the MHC class I genes HLA-B and HLA-A*. Nature, 2007. **450**(7171): p. 887-92.
51. Bergamaschi, L., et al., *HLA-class I markers and multiple sclerosis susceptibility in the Italian population*. Genes Immun, 2010. **11**(2): p. 173-80.
52. Bergamaschi, L., et al., *Association of HLA class I markers with multiple sclerosis in the Italian and UK population: evidence of two independent protective effects*. J Med Genet, 2011.
53. Ramagopalan, S.V. and G.C. Ebers, *Multiple sclerosis: major histocompatibility complexity and antigen presentation*. Genome medicine, 2009. **1**(11): p. 105.
54. Bitti, P.P., et al., *Association between the ancestral haplotype HLA A30B18DR3 and multiple sclerosis in central Sardinia*. Genetic epidemiology, 2001. **20**(2): p. 271-83.
55. Chao, M.J., et al., *Transmission of class I/II multi-locus MHC haplotypes and multiple sclerosis susceptibility: accounting for linkage disequilibrium*. Hum Mol Genet, 2007. **16**(16): p. 1951-8.
56. Healy, B.C., et al., *HLA B*44: protective effects in MS susceptibility and MRI outcome measures*. Neurology, 2010. **75**(7): p. 634-40.
57. Ramagopalan, S.V., et al., *Expression of the multiple sclerosis-associated MHC class II Allele HLA-DRB1*1501 is regulated by vitamin D*. PLoS genetics, 2009. **5**(2): p. e1000369.

58. Bayes, H.K., C.J. Weir, and C. O'Leary, *Timing of birth and risk of multiple sclerosis in the Scottish population*. European neurology, 2010. **63**(1): p. 36-40.
59. Staples, J., A.L. Ponsonby, and L. Lim, *Low maternal exposure to ultraviolet radiation in pregnancy, month of birth, and risk of multiple sclerosis in offspring: longitudinal analysis*. BMJ, 2010. **340**: p. c1640.
60. Templer, D.I., et al., *Season of birth in multiple sclerosis*. Acta Neurol Scand Suppl, 1992. **85**(2): p. 107-9.
61. Willer, C.J., et al., *Timing of birth and risk of multiple sclerosis: population based study*. BMJ, 2005. **330**(7483): p. 120.
62. Saastamoinen, K.P., M.K. Auvinen, and P.J. Tienari, *Month of birth is associated with multiple sclerosis but not with HLA-DR15 in Finland*. Multiple sclerosis, 2011.
63. Baarnhielm, M., et al., *Sunlight is associated with decreased multiple sclerosis risk: no interaction with human leukocyte antigen-DRB1*15*. European journal of neurology : the official journal of the European Federation of Neurological Societies, 2012.
64. Simpson, S., Jr., et al., *Latitude is significantly associated with the prevalence of multiple sclerosis: a meta-analysis*. Journal of neurology, neurosurgery, and psychiatry, 2011. **82**(10): p. 1132-41.
65. Buck, D., et al., *Influence of the HLA-DRB1 genotype on antibody development to interferon beta in multiple sclerosis*. Archives of neurology, 2011. **68**(4): p. 480-7.
66. Hoffmann, S., et al., *HLA-DRB1*0401 and HLA-DRB1*0408 are strongly associated with the development of antibodies against interferon-beta therapy in multiple sclerosis*. American journal of human genetics, 2008. **83**(2): p. 219-27.
67. Weber, F., et al., *Single-nucleotide polymorphisms in HLA- and non-HLA genes associated with the development of antibodies to interferon-beta therapy in multiple sclerosis patients*. The pharmacogenomics journal, 2011.
68. Sominanda, A., J. Hillert, and A. Fogdell-Hahn, *In vivo bioactivity of interferon-beta in multiple sclerosis patients with neutralising antibodies is titre-dependent*. Journal of neurology, neurosurgery, and psychiatry, 2008. **79**(1): p. 57-62.
69. Hegen, H., et al., *Persistency of neutralizing antibodies depends on titer and interferon-beta preparation*. Multiple sclerosis, 2011.
70. Sominanda, A., et al., *Inhibition of endogenous interferon beta by neutralizing antibodies against recombinant interferon beta*. Archives of neurology, 2010. **67**(9): p. 1095-101.
71. van der Voort, L.F., et al., *Clinical effect of neutralizing antibodies to interferon beta that persist long after cessation of therapy for multiple sclerosis*. Archives of neurology, 2010. **67**(4): p. 402-7.
72. Sominanda, A., et al., *Interferon beta preparations for the treatment of multiple sclerosis patients differ in neutralizing antibody seroprevalence and immunogenicity*. Multiple sclerosis, 2007. **13**(2): p. 208-14.
73. Handel, A.E., et al., *No evidence for an effect of DNA methylation on multiple sclerosis severity at HLA-DRB1*15 or HLA-DRB5*. Journal of neuroimmunology, 2010. **223**(1-2): p. 120-3.
74. Irizarry, R.A., et al., *The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores*. Nature genetics, 2009. **41**(2): p. 178-86.