Dissecting the Genetic Basis of Systemic Lupus Erythematosus

The Pursuit of Functional Variants

ANGÉLICA MARÍA DELGADO VEGA
Dissertation presented at Uppsala University to be publicly examined in Rudbecksalen, The Rudbeck Laboratory, Dag Hammarskjölds väg 20, Uppsala, Friday, April 26, 2013 at 09:15 for the degree of Doctor of Philosophy (Faculty of Medicine). The examination will be conducted in English.

Abstract

Systemic lupus erythematosus (SLE) is a chronic and systemic autoimmune disease that primarily affects women during the childbearing years. SLE is characterized by the production of autoantibodies against nucleic acids and their interacting proteins. The exact molecular mechanisms leading to the breakdown of self-tolerance remain to a large extent unknown, but it is well established that they are influenced by both non-genetic (i.e. environmental and hormonal) and genetic factors. SLE is a complex, polygenic disease. Several susceptibility variants have been identified in SLE. However, the functional role in disease pathogenesis for the majority of them remains largely unknown.

This thesis includes case-control association studies where the role of the genes TNFSF4 (Paper I), STAT4 (Paper II), CD226 (Paper III), and BLK (Papers IV and V) in the susceptibility of developing SLE was investigated. The primary focus was on the identification of the functional variants underlying the association. For each of these genes, fine mapping was performed using single nucleotide polymorphisms (SNPs), the linkage disequilibrium (LD) was characterized, and the association was narrowed down to specific haplotypes by means of several different statistical genetic strategies. Candidate variants were prioritized for further functional analysis on the basis of their potential effect on the gene function, their association, and/or biological plausibility. In Paper I, the association of TNFSF4 with SLE was validated and attributed to a risk haplotype tagged by SNPs rs1234317-T and rs12039904-T. Paper II provides evidence supporting the presence of at least two independent genetic effects within the STAT4 gene represented by rs3821236-A and rs7574865-A, which correlated with increased levels of gene expression. In Paper III, a functional allele in CD226 (rs727088-C) was identified, which was responsible for decreased levels in both mRNA and protein expression. In Paper IV, two independent genetic effects in the BLK gene were demonstrated. The first one comprised multiple regulatory variants in high LD that were enriched for NFκB and IRF4 binding sites and correlated with low BLK mRNA levels. The second was a low-frequency missense substitution (Ala71Thr) that decreased the BLK protein half-life. In Paper V, a genetic epistatic interaction between BANK1 rs10516487 (GG) and BLK rs2736340 (TT+TC) was demonstrated. Additional molecular analyses established that these molecules interact physically.

These studies have contributed to the dissection of the genetic architecture of SLE. They highlight the allelic heterogeneity of the disease and provide functional links to the associated variants, which has significantly aided in the understanding of SLE disease pathogenesis.

Keywords: Systemic Lupus Erythematosus, SLE, Genetic Mapping, Association Studies, Functional Variants, TNFSF4, STAT4, IRF5, CD226, BLK, BANK1

Angélica María Delgado Vega, Uppsala University, Department of Immunology, Genetics and Pathology, Medical Genetics, Rudbecklaboratoriet, SE-751 85 Uppsala, Sweden.

© Angélica Maria Delgado Vega 2013

ISSN 1651-6206
ISBN 978-91-554-8620-4
urn:nbn:se:uu:diva-196428 (http://urn.kb.se/resolve?urn=nbn:se:uu:diva-196428)
To Emi and T
Everything and anything for you
♥
List of Papers

This thesis is based on the following five papers, which are referred to in the text by their Roman numerals.


* Co-first authors

Reprints were made with permission from the respective publishers.
Related papers

Papers published during the PhD process, not included in this thesis.


* Co-first authors

No reprints were made.
# Contents

Introduction .............................................................................................................................................. 13
  Why is the work in this thesis important? ......................................................................................... 13
  Systemic Lupus Erythematosus as an Autoimmune Disease ....................................................... 14
    Autoimmune Diseases .................................................................................................................. 14
  Systemic Lupus Erythematosus ........................................................................................................ 16
  Systemic Lupus Erythematosus as a Complex Genetic Diseases ..................................................... 18
Finding the SLE Genes ......................................................................................................................... 19
  Linkage analysis ............................................................................................................................ 19
  Animal models ............................................................................................................................... 21
  Association studies ........................................................................................................................ 21
  Lessons from association studies of SLE ......................................................................................... 22
  The pursuit of functional variants and the “missing heritability” ..................................................... 30
  Non-genetic factors ....................................................................................................................... 34

The present investigation ...................................................................................................................... 35
  Aims .................................................................................................................................................. 35
Methods for Genetic Mapping of Disease Variants ............................................................................. 35
  Study subjects ............................................................................................................................... 35
  Selection of SNPs to be genotyped ............................................................................................... 36
  Genotyping ..................................................................................................................................... 36
  Statistical analysis ........................................................................................................................ 38

Results and Discussion ......................................................................................................................... 43
  Paper I: Replication of the TNFSF4 (OX40L) Association .......................................................... 43
  Paper II: Two Independent Effects in STAT4 that Correlate with Gene Expression and Act Additively with IRF5 .......................................................... 46
  Paper III: A functional 3'UTR Variant in CD226 Impairs Gene Expression in T and NKT Cells .............................................................................................................. 51
  Paper IV: Common and Rare Variants in BLK Lead to Reduced Half-Life of the Protein ............ 56
  Paper V: Statistical and Functional Epistasis of the B-cell Genes BANK1 and BLK .................... 60

Conclusions ............................................................................................................................................. 64

Future Perspectives ............................................................................................................................... 65

Resumen en Español ............................................................................................................................. 67
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>African-American</td>
</tr>
<tr>
<td>ADs</td>
<td>Autoimmune Disease(s)</td>
</tr>
<tr>
<td>AIMs</td>
<td>Ancestry informative markers</td>
</tr>
<tr>
<td>AITD</td>
<td>Autoimmune thyroid disease</td>
</tr>
<tr>
<td>AlAr</td>
<td>Alopecia Areata</td>
</tr>
<tr>
<td>ANAs</td>
<td>Antinuclear antibodies</td>
</tr>
<tr>
<td>Anti-dsDNA</td>
<td>Anti double stranded DNA antibodies</td>
</tr>
<tr>
<td>APCs</td>
<td>Antigen presenting cells</td>
</tr>
<tr>
<td>ASN</td>
<td>Asian</td>
</tr>
<tr>
<td>ASO</td>
<td>Allelic-specific oligonucleotides</td>
</tr>
<tr>
<td>ASP</td>
<td>Affected sibling pairs</td>
</tr>
<tr>
<td>BD</td>
<td>Behçet's disease</td>
</tr>
<tr>
<td>Bp</td>
<td>Bases pairs</td>
</tr>
<tr>
<td>CD-CV</td>
<td>Common disease - common variants hypothesis</td>
</tr>
<tr>
<td>CeD</td>
<td>Celiac disease</td>
</tr>
<tr>
<td>ChD</td>
<td>Crohn's disease</td>
</tr>
<tr>
<td>CNVs</td>
<td>Copy Number Variants</td>
</tr>
<tr>
<td>CTLs</td>
<td>Cytotoxic T lymphocytes</td>
</tr>
<tr>
<td>DCs</td>
<td>Dendritic cells</td>
</tr>
<tr>
<td>DILE</td>
<td>Drug-induced lupus erythematosus</td>
</tr>
<tr>
<td>EAMR</td>
<td>European-Amerindian</td>
</tr>
<tr>
<td>EBV</td>
<td>Epstein-Barr virus</td>
</tr>
<tr>
<td>EUR</td>
<td>European</td>
</tr>
<tr>
<td>FA</td>
<td>Familiar Autoimmunity</td>
</tr>
<tr>
<td>FRET</td>
<td>Fluorescence resonance energy transfer</td>
</tr>
<tr>
<td>GCA</td>
<td>Giant cell arteritis</td>
</tr>
<tr>
<td>GD</td>
<td>Grave's disease</td>
</tr>
<tr>
<td>GWAS</td>
<td>Genome-wide association studies</td>
</tr>
<tr>
<td>HLA</td>
<td>Human leukocyte antigens</td>
</tr>
<tr>
<td>HT</td>
<td>Hashimoto thyroiditis</td>
</tr>
<tr>
<td>HWE</td>
<td>Hardy-Weinberg equilibrium</td>
</tr>
<tr>
<td>IBD</td>
<td>Inflammatory bowel disease</td>
</tr>
<tr>
<td>IBS</td>
<td>Identical by state</td>
</tr>
<tr>
<td>IC</td>
<td>(Antibody-antigen) immune complexes</td>
</tr>
<tr>
<td>IFN</td>
<td>Interferon</td>
</tr>
<tr>
<td>INDEL</td>
<td>Insertion/deletion</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>JIA</td>
<td>Juvenile idiopathic arthritis</td>
</tr>
<tr>
<td>Kb</td>
<td>Kilobases (1000 base pairs)</td>
</tr>
<tr>
<td>KD</td>
<td>Kawasaki disease</td>
</tr>
<tr>
<td>KLKs</td>
<td>Kallikreins</td>
</tr>
<tr>
<td>LD</td>
<td>Linkage disequilibrium</td>
</tr>
<tr>
<td>LSO</td>
<td>Locus-specific oligonucleotides</td>
</tr>
<tr>
<td>MAF</td>
<td>Minor allele frequency</td>
</tr>
<tr>
<td>MAS</td>
<td>Multiple autoimmune syndrome</td>
</tr>
<tr>
<td>Mb</td>
<td>Megabases (one million base pairs)</td>
</tr>
<tr>
<td>MG</td>
<td>Myasthenia Gravis</td>
</tr>
<tr>
<td>MHC</td>
<td>Major Histocompatibility Complex</td>
</tr>
<tr>
<td>mPA</td>
<td>Microscopic polyangiitis</td>
</tr>
<tr>
<td>MS</td>
<td>Multiple Sclerosis</td>
</tr>
<tr>
<td>NK</td>
<td>Natural killer cells</td>
</tr>
<tr>
<td>NKT</td>
<td>Natural killer T cells</td>
</tr>
<tr>
<td>OR</td>
<td>Odds ratio</td>
</tr>
<tr>
<td>PAPS</td>
<td>Primary antiphospholipid syndrome</td>
</tr>
<tr>
<td>PBC</td>
<td>Primary biliary cirrhosis</td>
</tr>
<tr>
<td>PBMCs</td>
<td>Peripheral blood mononuclear cells</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>pDCs</td>
<td>Plasmacytoid dendritic cells</td>
</tr>
<tr>
<td>PS</td>
<td>Population stratification</td>
</tr>
<tr>
<td>pSS</td>
<td>Primary Sjögren's syndrome</td>
</tr>
<tr>
<td>RA</td>
<td>Rheumatoid Arthritis</td>
</tr>
<tr>
<td>SLE</td>
<td>Systemic Lupus Erythematosus</td>
</tr>
<tr>
<td>SNPs</td>
<td>Single nucleotide polymorphisms</td>
</tr>
<tr>
<td>SNVs</td>
<td>Single nucleotide variants</td>
</tr>
<tr>
<td>SSc</td>
<td>Systemic sclerosis</td>
</tr>
<tr>
<td>STRs</td>
<td>Short Tandem Repeats</td>
</tr>
<tr>
<td>T1D</td>
<td>Type 1 Diabetes</td>
</tr>
<tr>
<td>TFBSs</td>
<td>Transcription factor binding sites</td>
</tr>
<tr>
<td>UC</td>
<td>Ulcerative Colitis</td>
</tr>
<tr>
<td>WG</td>
<td>Wegener's granulomatosis</td>
</tr>
</tbody>
</table>
Introduction

Why is the work in this thesis important?

*Can you find a cure for lupus?* Is a question that often comes up after I explain to my relatives and friends what I do. Another question I had to answer came from a friend, who has severe lupus and participated as a study subject: *Did you find something in my genome that can be tested on my daughters to predict whether they are going to suffer from lupus?* My answer to both questions was unfortunately: no, we cannot, not yet; and I tried to explain why. I have many times wondered when we will be able to answer these questions positively. I am sure many human geneticists have to answer these questions daily, to their patients, to funding agencies, to the media. In general, to a society that is awaiting, full of high expectations, to see the results of genetic research be reflected in better health care alternatives, sensitive and accurate prediction and diagnostic methods, and more efficient or less harmful treatments, at least for their children.

A vast wealth of knowledge about the genetic architecture of normal traits and diseases has been generated in the last decade, after the initial draft of the human genome was published (1, 2). The sequencing of the human genome provided a fundamental resource for biomedical research, as it made possible to develop a systematic, comprehensive and unbiased way to search for genes underlying disease. Most importantly, it has significantly deepened our understanding of human biology, evolution and history, and the biology of the human genome itself.

After ten years, genome-wide approaches had led to the identification of ~2,850 genes underlying monogenic disorders, ~1,100 genes underlying complex traits and ~150 targets of somatic mutation in cancer (3). Despite this undeniable success, the promise of genomics for medicine remains to be fully realized. Geneticists have the responsibility to openly deliver the results of genomic research to society while being clear about its limitations as well as its potential applications. The main medical potential of genomics relies on its ability to aid the understanding of the mechanisms underlying disease by revealing the pieces of the puzzle, which can only be built by integrating the information generated by other disciplines like biochemistry, proteomics, animal studies, among others. As a consequence, genomics may guide the development of new therapies because associated genes point to the cellular
pathways that underlie disease pathogenesis and, therefore, provides novel candidates to be therapeutically targeted.

Although many genes have been reliably associated with systemic lupus erythematosus (SLE) and other autoimmune diseases (ADs), the functional variants for the majority of the associated genes have not yet been found. For that reason, the biology underlying them is largely unknown. It is not sufficient to provide a list of associated variants or pieces of the puzzle; we need to understand how they function, how they interact with each other, and what happens if one piece is missing or overrepresented. Once we begin to understand this, we might then be able to improve therapies. However, to be able to predict which individuals will develop a disease we need to know how many pieces of the puzzle we have to look for, that is, how many genes and non-genetic factors are involved. For polygenic diseases like SLE that remains largely unknown. To date, the genetic variants associated with SLE explain only a small proportion of the familial aggregation. In addition, only a handful of non-genetic factors, environmental and hormonal, which also influence the risk to develop the disease have been identified.

This doctoral thesis aimed to dissect the association of four genes associated with SLE: *TNFSF4*, *STAT4*, *CD226* and *BLK*. In general, for these genes, I performed fine mapping using single nucleotide polymorphisms (SNPs), characterized the linkage disequilibrium (LD), and narrowed down the association to specific haplotypes by means of different statistical genetic strategies including SNP window analysis, conditional tests, imputation and interaction analysis. Candidate variants were prioritized for further functional analysis on the basis of their potential effect on the gene function, their association and/or biological plausibility. An overview of the genetic basis of SLE and actual challenges is provided as an introduction.

**Systemic Lupus Erythematosus as an Autoimmune Disease**

**Autoimmune Diseases**

The human immune system is able to sense and neutralize any possible molecule entering the body and usually recognizes but does not react to components of our own body. This is achieved thanks to a series of cellular checkpoints that keep in control the activity induced by the receptors that recognize self-antigens (4). The specific mechanisms underlying this self-tolerance and how defects in these controls lead to ADs are still undefined (5). While *autoimmunity* refers to physiological responses that do not lead to immunopathology, an *autoimmune disease* is a clinical syndrome caused by
the breakdown of immune tolerance and activation of self-reactive T cells and/or B cells, which in turn leads to histopathological damage of tissues and organs, in the absence of an infection, cancer or other discernible cause (6).

ADs are common and painful diseases. Although it is often stated that individual ADs are rare, as a group they are not. Their incidence and prevalence vary widely depending on the disorder and the ethnic background of the population. In general, taking them as a group, they have an incidence of 90 per 100,000 people per year and a prevalence that ranges between 3 and 5% with 85% of the cases being women (7). They are often severe, lifelong diseases, with a considerable burden on the quality of life of the individuals who suffer them, on their families and on the health system. Moreover, they are among the leading causes of death among young and middle-aged women in the U.S (8).

ADs comprise a group of at least hundred different although at the same time similar diseases. Despite the diversity in clinical presentation and affected organs, they seem to have common pathogenic mechanisms. Often several members of the same family can be affected with different ADs, and it is frequent to find multiple ADs co-occurring in the same patient. These examples are better known as familial autoimmunity (FA) and multiple autoimmune syndrome (MAS), respectively (9, 10). In addition, ADs share both non-genetic and genetic etiologic factors. Many susceptibility genes are shared by several ADs, whereas a few are disease-specific (11). All these findings support the idea of a common origin for ADs.

ADs can be either organ-specific or systemic. The organ-specific ADs are characterized by a cellular and/or humoral autoimmune response against one specific organ or system and localized histopathological damage. For example, the pancreas in type 1 diabetes (T1D), the central nervous system in multiple sclerosis (MS), and the thyroid gland in the Hashimoto thyroiditis (HT). In contrast, systemic autoimmune diseases are characterized by the production of non-organ specific autoantibodies and the involvement of multiple organs and systems; for example in SLE, rheumatoid arthritis (RA) and systemic sclerosis (SSc) (reviewed in (12)). There are also acute ADs (i.e. Henoch-Schonlein purpura and subacute thyroiditis), which usually display complete remission after a couple of weeks of treatment. On the contrary, chronic ADs are lifelong diseases characterized by alternate periods of partial remissions and exacerbations or “flares” of the disease activity. The chronic and intermittent inflammatory damage accumulates over time leading to the irreversible destruction of the respective target organs.
Systemic Lupus Erythematosus

SLE (MIM 152700) is an often-severe chronic and systemic AD characterized by loss of immune tolerance to multiple cellular components, especially nucleic acids and their interacting proteins. SLE affects women in 90% of the cases with a peak of incidence during childbearing years (15–44 years of age), thus suggesting that hormonal factors may trigger the disease onset and flares (13). The prevalence of SLE ranges from around 20 cases per 100,000 people in Europeans to more than 200 cases per 100,000 individuals with African ancestry (13, 14). In Sweden, the prevalence of SLE is around 40 cases per 100,000 individuals. These broad differences in disease prevalence among populations may be attributed to different genetic backgrounds or differences in local environmental factors (i.e. diet, UV exposition).

The clinical presentation of SLE is very heterogeneous. The classification of SLE disease status is determined by meeting at least four out of eleven criteria (Table 1)(15) that range from rashes through arthritis and nephritis to psychosis. Thus, the clinical presentation of the disease in terms of signs, symptoms, and severity might vary considerably between individuals and even in the same patient over time. At the serological level, SLE is characterized by the production of autoantibodies against nuclear targets, with antinuclear antibodies (ANAs) being the most common type present in 95% of patients. Antibodies against different nucleosome components (anti-double stranded DNA –dsDNA- and anti-histones), spliceosome (anti-Sm and anti-U1 RNP), cytoplasm ribonucleoproteins (anti-Ro/SSA, anti-La/SSB), and membrane phospholipids (i.e. anticardiolipin, antiprothrombin) have also been identified. The histopathological damage of SLE is driven by the autoantibodies, either through the formation and deposition of autoantibody–autoantigen immune complexes in multiple organs and tissues or through direct targeting to basement membranes, especially the glomerular basement membrane (GBM), which causes glomerulonephritis. The presence of different antibodies may predict different clinical outcomes (16). The ubiquity of the auto antigens explains why SLE can affect virtually any organ Therefore, SLE is known as the prototype of systemic ADs. Patients with SLE also suffer of manifestations secondary to treatment toxicity, like infections caused by the immunosuppressive medication that is used to reduce the immune system hyperactivity in an attempt to control the irreversible organ damage. Infections and renal failure are among the most common causes of mortality in young SLE patients (17).
<table>
<thead>
<tr>
<th>Criterion</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Malar Rash</td>
<td>Fixed erythema, flat or raised, over the malar eminences, tending to spare the nasolabial folds</td>
</tr>
<tr>
<td>2. Discoid rash</td>
<td>Erythematous raised patches with adherent keratotic scaling and follicular plugging; atrophic scarring may occur in older lesions</td>
</tr>
<tr>
<td>3. Photosensitivity</td>
<td>Skin rash as a result of unusual reaction to sunlight, by patient history or physician observation</td>
</tr>
<tr>
<td>4. Oral ulcers</td>
<td>Oral or nasopharyngeal ulceration, usually painless, observed by physician</td>
</tr>
<tr>
<td>5. Nonerosive Arthritis</td>
<td>Involving 2 or more peripheral joints, characterized by tenderness, swelling, or effusion</td>
</tr>
<tr>
<td>6. Pleuritis or Pericarditis</td>
<td>Pleuritis--convincing history of pleuritic pain or rubbing heard by a physician or evidence of pleural effusion OR Pericarditis--documented by electrocardiogram or rub or evidence of pericardial effusion</td>
</tr>
<tr>
<td>7. Renal Disorder</td>
<td>Persistent proteinuria &gt; 0.5 grams per day or &gt; than 3+ if quantification not performed OR Cellular casts--may be red cell, hemoglobin, granular, tubular, or mixed</td>
</tr>
<tr>
<td>8. Neurologic Disorder</td>
<td>Seizures--in the absence of offending drugs or known metabolic derangements; e.g., uremia, ketoacidosis, or electrolyte imbalance OR Psychosis--in the absence of offending drugs or known metabolic derangements, e.g., uremia, ketoacidosis, or electrolyte imbalance</td>
</tr>
<tr>
<td>9. Hematologic Disorder</td>
<td>Hemolytic anemia--with reticulocytosis OR Leucopenia--&lt; 4,000/mm³ on ≥ 2 occasions OR Lymphopenia--&lt; 1,500/ mm³ on ≥ 2 occasions OR Thrombocytopenia--&lt;100,000/ mm³ in the absence of offending drugs</td>
</tr>
<tr>
<td>10. Immunologic Disorder</td>
<td>Anti-DNA: antibody to native DNA in abnormal titer OR Anti-Sm: presence of antibody to Sm nuclear antigen OR Positive finding of antiphospholipid antibodies on: 1. An abnormal serum level of IgG or IgM antcardiolipin antibodies, 2. A positive test result for lupus anticoagulant using a standard method, or 3. A false-positive test result for at least 6 months confirmed by Treponema pallidum immobilization or fluorescent treponemal antibody absorption test</td>
</tr>
<tr>
<td>11. Positive Antinuclear Antibody</td>
<td>An abnormal titer of antinuclear antibody by immunofluorescence or an equivalent assay at any point in time and in the absence of drugs</td>
</tr>
</tbody>
</table>
The pathogenic events behind SLE have not yet been completely elucidated. However, the evidence available today has allowed the formulation of a model in which environmental and hormonal exposures in a genetically susceptible individual may lead to the abnormal exposition to auto-antigens and aberrant activation of both innate and adaptive immune system, which results in self-tolerance breakdown, production and deposition of auto-antibodies and cytokines that mediate inflammation and tissue injury. The pre-clinical phase can be long and once the disease is expressed clinically organ damage may have already occurred. The immune amplification leads to further inflammation and often ends in irreversible organ damage (18, 19). The identification of the genetic, epigenetic, environmental, and hormonal factors that contribute to the disease process are starting to lead to the development of better classification of patients ad more rational treatments of the disease (18, 19).

Systemic Lupus Erythematosus as a Complex Genetic Diseases

From a genetic point of view, SLE is a complex genetic disease meaning that its inheritance does not follow a Mendelian-like model and it is polygenic. However, in rare cases, single gene defects may lead to the development of SLE, for example the homozygous deficiency of classical complement pathway genes (C1Q, C1R, C1S, C4A) (20). The first suggestion of a genetic component in SLE came from the evidence of familial aggregation. The prevalence of the disease is higher in siblings of patients compared to the general population (λs=8–29) (21), the concordance rates between monozygotic twins (20-40%) is about ten times higher relative to dizygotic twins and other full siblings (2–5%) (22). These studies provided the first lines of evidence for a genetic component in SLE and justified the search for susceptibility genes.

Early studies on segregation analysis predicted that the identification of the genetic variation leading to increased susceptibility to SLE was going to be an arduous task. They postulated a probable model of inheritance in which multiple minor polygenic effects were acting in an additive fashion and estimated that the heritability of the disease was around 66% (23, 24). Thus, non-genetic factors also play an important role. The methods used to map disease genes have developed rapidly, thanks to the increasing knowledge about the human genome, which has been closely linked to the rapid development of new high-throughput genotyping technologies and statistical methods that have made it possible to study the genetic diversity among individuals and populations at an increasing number of loci at a reasonable
cost. The papers included in this thesis reflect the shift in technology and the available public resources that have occurred over a short period of time, making possible more comprehensive genetic studies.

Finding the SLE Genes

Genetic mapping refers to the systemic search for genotypic variation that correlates with phenotypic variation across the genome, without any previous hypothesis about the localization of the DNA variants. In order to identify the genetic loci (physical sites on the chromosome) contributing to disease susceptibility two main genetic mapping strategies have been followed: genome-wide linkage analysis and association studies. Both strategies take advantage of the most common types of DNA variation present in the human genome, summarized in Figure 1.

Linkage analysis

Linkage analysis aims to identify genomic regions containing susceptibility genes by studying the co-segregation of genetic variants, primarily microsatellites, through a family in conjunction with the disease trait. In complex diseases like SLE, linkage studies have used non-parametric methods in affected sibling pairs (ASP), based on the assumption that ASPs will exhibit excess sharing of haplotypes identical by state (IBS) in the genomic regions where susceptibility genes are located. Once a linkage region is identified, the aim is to narrow it down until the disease variants are identified. This can be achieved either by further fine mapping (positional approach), or by choosing the best candidate genes of the linkage region based on previous knowledge about their function, expression pattern, phenotypes in knock out or transgenic animal models, or association with other ADs (biological plausibility). Historically, linkage studies had partial success in the dissection of susceptibility genes for SLE, given their limited power to detect genes with small to moderate effect, and their sensitivity to genetic and clinical heterogeneity. However, linkage analysis and/or association studies of candidate genes in linkage regions have contributed to the identification of some SLE susceptibility genes. For example, HLA-DRB1 and TNF (tumour necrosis factor) in the major histocompatibility complex (MHC), which is the strongest and most consistent genomic region linked to SLE (25), the immunoglobulin Fcγ receptors FCGR2A and FCGR3A on 1q23 (26), TNFSF4 on 1q25 (27, 28), PDCD1 on 2q37 (29), and the Kallikrein gene cluster (KLKs) on 19q13 (30). Linkage analyses also provided evidence that the polygenic additive model was rather simplistic and that the complexity of SLE as a genetic disease is further complicated by incomplete penetrance, genetic heterogeneity, and allelic heterogeneity.
Figure 1. DNA variants commonly used for genetic mapping of human diseases.
The figure shows five stretches of DNA sequences on the same chromosome with different types of common DNA variation. The major alleles are shown in blue and the minor alleles in red. Single nucleotide polymorphisms (SNPs) are the most abundant type of genetic variation (~1 of every 300 nucleotides) and are able to define ancestral chromosome segments shared across populations. SNPs are considered common if the minor allele frequency (MAF) is >5%, low-frequency if the MAF is between 1-5% and rare if the MAF is <1%. An insertion/deletion polymorphism (INDEL) occurs when one or two non-repeated nucleotides are inserted or deleted. Microsatellites or short tandem repeat (STR) units of 1, 2 or 4 nucleotides are also common. STRs can have from 5 to 20 repeat units (alleles) and are thus highly informative. It is also common to find variable copies of DNA sequences of at least 1kb in length (copy number variants or CNVs). They usually form tandem clusters and are responsible for the greatest percentage of nucleotides differing between two individuals. The figure also illustrates how the alleles of SNP1, SNP2 and the INDEL are perfectly correlated or in total linkage disequilibrium (LD). As a consequence, only a few allele combinations (haplotypes) of the $2^3$ possible are observed. SNP3, SNP4 and the STR are also in perfect LD. There is no correlation between the two haplotype blocks as they are separated by a recombination hotspot. In the lower panel, a square represents the LD ($D'$) between each pair of SNPs and the intensity of LD is represented in red.
Animal models

Linkage analyses in murine models that develop disorders resembling SLE have identified several loci where the homologous genes in humans are obvious candidate genes for human SLE. This strategy has led to the identification of several SLE genes including FCGR2B, SLAMF3 (LY9)/SLAMF4 (CD244), CR2, KLK genes, FAS and TLR7, which are the human homologous of Fcgr2b, Sle1b (Ly108), Sle1c1 (cr2), and Sle3 in NZM2410 mice; Fas lpr in MRL/lpr mice and Yaa/tlr7 in BXSB/Yaa mice, respectively (30-42). Studies in a canine model of SLE (the Nova Scotia duck tolling retriever breed) have identified several genomic regions associated with SLE containing multiple genes (43). The high LD characteristic of this dog breed makes it difficult to determine which are the genes underlying disease. However, these regions contain interesting candidates involved in T cell and B cell activity, including BANK1, which is a human SLE susceptibility locus (44).

Association studies

Association studies map susceptibility effects through the detection of statistical differences in allele and genotype frequencies between individuals with a trait of interest and a set of carefully matched controls, which are healthy subjects of the same gender, age, and ancestry. Unlike linkage, which points to a chromosome marker independently of the allele, association describes the connection of the trait to a specific allele and/or the haplotype tagged by it. Initially, association studies were based on candidate genes. Later, the initial phase of The International HapMap Project made possible the design and analysis of genome-wide association studies (GWAS) since it provided a catalogue of ~1 million single nucleotide polymorphism (SNPs) in Caucasian, Asian and African human populations (45). HapMap also revealed that the alleles of SNPs located nearby are usually in high linkage disequilibrium (LD) and form haplotype blocks of 7.3 to 16.3 kb, depending on the population. The blocks are characterized by low haplotype diversity and are separated by recombination hotspots (Figure 1)(45). The catalogue grew quickly and by 2007 it contained ~10 million common variants from The Human Genome Project (1), the SNP Consortium (46) and the International HapMap Project (47). A catalogue of copy number variants (CNVs) (Figure 1) was also generated (48, 49). Despite being less numerous than SNPs, CNVs cover ~3.5 Mb (~0.1%) of the human genome, about one-third overlap coding regions, and together they are responsible for the greatest number of nucleotides that differ between any two individuals (49). The latest versions of GWAS arrays have also integrated SNPs tagging CNVs.
The pattern of LD across the human genome is shaped by the demographic history of each population. For that reason, there was controversy around the transferability of LD patterns to populations not included in HapMap (50, 51). However, several studies supported that a minimum set of ~500,000 SNPs could capture most of the common genetic variants across populations (90 to 95% of the alleles with a MAF > 5%) if they were highly correlated with the non-typed SNPs ($r^2$>0.80) (50, 51). The easy access and interpretation of the data generated by HapMap made SNPs the preferred type of polymorphisms for association studies, in combination with the rapid development of relatively cheap and high-throughput SNP genotyping arrays.

**Lessons from association studies of SLE**

The success of GWAS in the identification of susceptibility genes for common diseases is undeniable. Around 400 variants pointing to ~100 loci have been associated with ADs and around half of them are shared by at least two ADs (52, 53) ([Figure 2](#) and **Figure 3**). As of January 8th 2013, ten GWAS in SLE (44, 54-62) were registered in the NHGRI GWAS catalogue ([http://www.genome.gov/gwastudies/](http://www.genome.gov/gwastudies/)). The have successfully identified ~50 susceptibility loci in European and Asians populations, with some loci being population-specific. Several large replication studies and meta-analysis of association studies have increased the list of SLE susceptibility genes (Table 2). Although GWAS are yet to be performed in SLE in other populations, the European and Asian associated variants have been tested in African-Americans (63) and European-Amerindians (64). While the association of many of the variants have been replicated, the allelic frequencies vary widely between populations and, consequently, the relative role of the susceptibility genes may differ. For example, the frequency of European risk variants is in general significantly higher in European-Amerindians (64), while $PTPN22\ rs2476601$ and $PDCD1\ rs11568821$ risk alleles are rare in African-Americans and absent in Asians (65, 66). The combination of genes that confer susceptibility to SLE varies across individuals, families and populations, and thus the genetic heterogeneity of the disease is very high.
Figure 2. Non-MHC shared loci between GWAS of autoimmune diseases. For every associated locus, the reported gene(s) and the trait are displayed below and above the chromosome, respectively. Copyright © 2011 Ramos et al. (53)
Figure 3. MHC shared association between GWAS of autoimmune diseases. The MHC spans ~7Mb on the short arm of chromosome 6 (6p21) and contains over 400 gene, divided intro three classes. The class II genes HLA-DRB1 and HLA-DQB1 and are the strongest genetic susceptibility factors to SLE. Figure made using the gwascat R package version 1.2.1 by VJ Carey.
Table 2. Genes associated with susceptibility to develop SLE

<table>
<thead>
<tr>
<th>Chr</th>
<th>Gene</th>
<th>SLE pathway/function</th>
<th>Population</th>
<th>Other ADs</th>
<th>Association with ACR criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>1p13</td>
<td>PTPN22</td>
<td>Modulation of lymphocyte receptor activation</td>
<td>EUR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1p31.1</td>
<td>NEGR1</td>
<td></td>
<td>EUR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1p32.3</td>
<td>SLC1A7</td>
<td></td>
<td>EUR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1q23</td>
<td>FCGR3B</td>
<td>Clearance of ICs</td>
<td>EUR, AA, EAMR</td>
<td>mPA, WG</td>
<td>Renal</td>
</tr>
<tr>
<td>1q23</td>
<td>FCGR2A</td>
<td>Phagocytosis and clearance of ICs</td>
<td>EUR, AA, EAMR</td>
<td>PAPS</td>
<td>Malar rash, renal</td>
</tr>
<tr>
<td>1q23</td>
<td>FCGR3A</td>
<td>Phagocytosis and clearance of ICs</td>
<td>EUR</td>
<td></td>
<td>Lupus nephritis</td>
</tr>
<tr>
<td>1q23</td>
<td>FCGR2B</td>
<td>Phagocytosis of ICs and regulation of antibody production by B-cells</td>
<td>EUR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1q24</td>
<td>IL10</td>
<td>Inhibits T cells and APCs, enhances B cell survival and activity.</td>
<td>EUR</td>
<td></td>
<td>Anti-RBP, discoid rash, neurological, renal</td>
</tr>
<tr>
<td>1q25.1</td>
<td>TNFSF4</td>
<td>T cell-APC interaction</td>
<td>EUR, ASN, EAMR</td>
<td>CeD, ChD, MS, pSS</td>
<td></td>
</tr>
<tr>
<td>1q25.3</td>
<td>LAMC2</td>
<td></td>
<td>EUR</td>
<td></td>
<td>Anti-dsDNA (+)</td>
</tr>
<tr>
<td>2p22.3</td>
<td>RASGRP3</td>
<td>BCR signalling via Ras-ERK</td>
<td>ASN</td>
<td></td>
<td>ANAs, malar rash, discoid rash</td>
</tr>
<tr>
<td>2q24</td>
<td>IFIH1</td>
<td>Cytoplasmic sensor of nucleic acids</td>
<td>EUR</td>
<td></td>
<td>Anti-dsDNA (+)</td>
</tr>
<tr>
<td>2q32.3</td>
<td>STAT4</td>
<td>Signalling of IL-12, IL-23 or IFNα/β, production of IFNγ, Th cells development</td>
<td>EUR, ASN, AA, EAMR</td>
<td>RA, PBC, SS, BD, CeD, Psoriasis, pSS, PAPS</td>
<td>Anti-dsDNA (+), oral ulcers, renal, early onset</td>
</tr>
<tr>
<td>2q37</td>
<td>PDCD1</td>
<td>Regulation of T cell activity</td>
<td>EUR, ASN, EAMR</td>
<td>RA</td>
<td></td>
</tr>
<tr>
<td>3p14.3</td>
<td>PXK</td>
<td></td>
<td>EUR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3p21.3</td>
<td>TREX1</td>
<td>Exonuclease involved in cell death, DNA degradation and cellular response to oxidative damage</td>
<td>EUR</td>
<td>pSS</td>
<td>Neurological, ANAs</td>
</tr>
<tr>
<td>3q26.31</td>
<td>NAAALDL2</td>
<td></td>
<td>EUR</td>
<td>KD</td>
<td>Anti-dsDNA(+)</td>
</tr>
<tr>
<td>4q21.3</td>
<td>AFF1</td>
<td>B Lymphocyte development;</td>
<td>ASN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4q22.1</td>
<td>SPP1</td>
<td>Production of IFN-α</td>
<td>EUR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4q24</td>
<td>BANK1</td>
<td>BCR-induced calcium mobilization. Scaffold of Src kinases</td>
<td>EUR, ASN, EAMR</td>
<td></td>
<td>Anti-dsDNA(+)</td>
</tr>
<tr>
<td>4q25</td>
<td>LEFI</td>
<td></td>
<td>ASN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4q25</td>
<td>COL25A1</td>
<td></td>
<td>EUR</td>
<td></td>
<td>Anti-dsDNA(+)</td>
</tr>
<tr>
<td>5p12</td>
<td>GHR</td>
<td></td>
<td>EUR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chromosome</td>
<td>Gene</td>
<td>Description</td>
<td>Population(s)</td>
<td>Associated Conditions</td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>------</td>
<td>-------------</td>
<td>----------------</td>
<td>-----------------------</td>
<td></td>
</tr>
<tr>
<td>5q33.1</td>
<td><strong>TNIP1</strong></td>
<td>Regulation of NFκB activation</td>
<td>EUR, ASN</td>
<td>SSc, MG, Psoriasis, Vasculitis</td>
<td></td>
</tr>
<tr>
<td>5q33.3</td>
<td><strong>PTTG1</strong></td>
<td></td>
<td>EUR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5q34</td>
<td><strong>MIR146A</strong></td>
<td>Negative regulator of IFN pathway</td>
<td>EUR, ASN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6p21.31</td>
<td><strong>UHRF1BP1</strong></td>
<td></td>
<td>EUR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6p21.32</td>
<td><strong>HLA-DRB1</strong></td>
<td>Antigen presentation</td>
<td>EUR, ASN, AA, EAMR</td>
<td>JIA, MS, RA, SSc, TID, UC, Anti-RBP, anti-dsDNA(+)</td>
<td></td>
</tr>
<tr>
<td>6p21.32</td>
<td><strong>HLA-DQA1</strong></td>
<td>Antigen presentation</td>
<td>EUR, ASN</td>
<td>UC, Vitiligo, IBD, ChD, RA, CeD, SSc, Anti-dsDNA(+)</td>
<td></td>
</tr>
<tr>
<td>6p21.32</td>
<td><strong>HLA-DQA2</strong></td>
<td>Antigen presentation</td>
<td>EUR</td>
<td>RA, AIAr, ChD Anti-dsDNA(-)</td>
<td></td>
</tr>
<tr>
<td>6p21.32</td>
<td><strong>NOTCH4</strong></td>
<td></td>
<td>EUR</td>
<td>SSc, Anti-dsDNA(+)</td>
<td></td>
</tr>
<tr>
<td>6p21.33</td>
<td><strong>MSH5</strong></td>
<td></td>
<td>EUR, EAMR</td>
<td>Anti-dsDNA(+)</td>
<td></td>
</tr>
<tr>
<td>6p21.33</td>
<td><strong>PRDM1</strong></td>
<td>IFN-β expression, regulator B-cell and T-cell differentiation.</td>
<td>ASN</td>
<td>ChD, Anti-dsDNA (+), Anti-RBP</td>
<td></td>
</tr>
<tr>
<td>6q21</td>
<td><strong>ATG5</strong></td>
<td>Autophagocytosis, caspase-dependent apoptosis from FAS and TNF ligands</td>
<td>EUR, ASN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6q23.3</td>
<td><strong>TNFAIP3</strong></td>
<td>Inhibit NFκB activation and TNF-mediated apoptosis</td>
<td>EUR, ASN, AA, EAMR</td>
<td>RA, Psoriasis, SSc, Anti-dsDNA(+) and Anti-dsDNA(-), Renal, hematological</td>
<td></td>
</tr>
<tr>
<td>7p12.2</td>
<td><strong>IKZF1</strong></td>
<td>Lymphocyte differentiation, proliferation and BCR signalling</td>
<td>ASN, EAMR</td>
<td>CeD, Malar rash, renal</td>
<td></td>
</tr>
<tr>
<td>7q32.1</td>
<td><strong>IRF5</strong></td>
<td>Production of type 1 IFN and inflammatory cytokines upon TLR signalling</td>
<td>EUR, ASN, AA, EAMR</td>
<td>SSc, RA, PBC, Anti-dsDNA(+) and UC, pSS, TID (-), Anti-RBP</td>
<td></td>
</tr>
<tr>
<td>7q32.1</td>
<td><strong>TNPO3</strong></td>
<td></td>
<td>EUR</td>
<td>PBC, UC, SSc</td>
<td></td>
</tr>
<tr>
<td>8p23.1</td>
<td><strong>BLK</strong></td>
<td>BCR signalling and B-cell development</td>
<td>EUR, ASN</td>
<td>RA, KD, PAPS, Anti-dsDNA (+), SSc, pSS</td>
<td></td>
</tr>
<tr>
<td>8p23.1</td>
<td><strong>XKR6</strong></td>
<td></td>
<td>EUR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8q13</td>
<td><strong>LYN</strong></td>
<td>BCR signalling</td>
<td>EUR</td>
<td>Discoid rash, haematological</td>
<td></td>
</tr>
<tr>
<td>9q34.13</td>
<td><strong>NTNG2</strong></td>
<td></td>
<td>EUR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10q11.22</td>
<td><strong>LRRC18</strong></td>
<td></td>
<td>ASN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10q11.23</td>
<td><strong>WDFY4</strong></td>
<td></td>
<td>ASN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11p15.5</td>
<td><strong>KIAA1542</strong></td>
<td></td>
<td>EUR</td>
<td>Anti-dsDNA (+)</td>
<td></td>
</tr>
<tr>
<td>11p15.5</td>
<td><strong>IRF7</strong></td>
<td>Type I IFN pathway</td>
<td>EUR, ASN, EAM</td>
<td>Anti-dsDNA, anti-Sm</td>
<td></td>
</tr>
<tr>
<td>Chr</td>
<td>Gene</td>
<td>Function</td>
<td>Ethnicity</td>
<td>Other Conditions</td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>------</td>
<td>----------</td>
<td>-----------</td>
<td>-----------------</td>
<td></td>
</tr>
<tr>
<td>11q11</td>
<td>OR4A15</td>
<td>B-cell and T helper type 17 cell differentiation</td>
<td>EUR</td>
<td>Anti-dsDNA (+)</td>
<td></td>
</tr>
<tr>
<td>11q24.3</td>
<td>ETS1</td>
<td>Activates the LYN and BLK promoters</td>
<td>ASN</td>
<td>CeD, RA</td>
<td>Early onset</td>
</tr>
<tr>
<td>12q21.2</td>
<td>RPL7AP59</td>
<td>NFκB inflammatory pathway</td>
<td>EUR</td>
<td>Anti-dsDNA (-)</td>
<td></td>
</tr>
<tr>
<td>12q24.32SLC15A4</td>
<td></td>
<td>Complement coated particles and ICs clearance; leukocyte adhesion</td>
<td>ASN</td>
<td>Discoid rash</td>
<td></td>
</tr>
<tr>
<td>13q14.11ELF1</td>
<td></td>
<td>Complement coated particles and ICs clearance; leukocyte adhesion</td>
<td>EUR, ASN, AA, EAMR</td>
<td>Anti-dsDNA (+), discoid rash, renal, haematological</td>
<td></td>
</tr>
<tr>
<td>16p11.2</td>
<td>ITGAM</td>
<td>Complement coated particles and ICs clearance; leukocyte adhesion</td>
<td>EUR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16p11.2</td>
<td>ITGAX</td>
<td>Complement coated particles and ICs clearance; leukocyte adhesion</td>
<td>EUR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16p13.3</td>
<td>DNASE1</td>
<td>DNA degradation</td>
<td>AITD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16q24.1</td>
<td>IRF8</td>
<td>Type I IFN pathway</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18q22.3</td>
<td>SOCS6</td>
<td>Suppressor of cytokine signalling</td>
<td>EUR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18q22.3</td>
<td>CD226</td>
<td>T cell development, anti-apoptotic, NKT cell cytotoxic activity</td>
<td>EUR, EAMR, T1D, RA, MS, WG, CeD, AITD, SSc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19p13.2</td>
<td>TYK2</td>
<td>Response to type I IFN, phosphorylation of STATs</td>
<td>EUR</td>
<td>MS, Psoriasis</td>
<td></td>
</tr>
<tr>
<td>20p13</td>
<td>TMC2</td>
<td></td>
<td>EUR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22q11.21UBE2L3</td>
<td></td>
<td>Ubiquitination of NFκB precursors</td>
<td>EUR, ASN</td>
<td>CeD, RA</td>
<td>Anti-dsDNA (+)</td>
</tr>
<tr>
<td>22q11.21HIC2</td>
<td></td>
<td></td>
<td>ASN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xp22.3</td>
<td>TLR7</td>
<td>Pathogen recognition and activation of innate immunity</td>
<td>EUR, ASN, EAMR</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Chr: chromosome; EUR: European; ASN: Asian; AA: African-American; EAMR: European-Amerindian; ICs: immune complexes; anti-RBP: antibodies anti ribosomal binding proteins (Sm/RNP, SSA/SSB); anti-dsDNA: antibodies anti double stranded DNA; ANAs: antinuclear antibodies; SLE: systemic lupus erythematosus; PAPS: primary anti-phospholipid syndrome; T1D: type 1 diabetes; RA: rheumatoid arthritis; MS: multiple sclerosis; GD: Graves' disease; IBD: inflammatory bowel disease; pSS: primary Sjögren's syndrome; mPA: microscopic polyangiitis; WG: Wegener's granulomatosis; AITD: autoimmune thyroid diseases; JIA: juvenile idiopathic arthritis; GCA: giantcell arteritis; UC: ulcerative colitis; ChD: Crohn's disease; CeD: celiac disease; PBC: primary biliary cirrhosis; SSc: systemic sclerosis; BD: Behçet's disease; KD: Kawasaki disease; MG: Myasthenia Gravis; AlAr: Alopecia Areata;
The effect size of significant associations is often presented as the odds ratio (OR), its respective 95% confidence interval, and P-value. For a hypothetical SNP with two alternative alleles \(A\) and \(a\), an OR=3 for allele \(A\) means that the odds of having the disease is three times higher if carrying allele \(A\) than the odds if carrying allele \(a\). An OR \(\approx 1.0\) indicates no relationship between the allele and the case status, alleles with ORs < 1.0 are usually referred as “protective”, and those with ORs > 1.0 as “risk” alleles. The effect sizes of the SLE associated variants are low to moderate with ORs ranging between 1.1 and 1.8, with the notable exception of MHC loci. As a consequence, large collections of patients and controls in the order of several thousands are required to achieve the statistical power to detect effects of low/moderate size at the stringent statistical thresholds defined. If we take each SNP in a GWAS as an independent hypothesis test and consider that hundreds of thousands to millions of SNPs are tested, the probability of finding significant associations just by chance (type I error) is very high. Therefore, corrections for multiple testing and stringent significance thresholds to avoid false positives are needed. A P-value < \(1 \times 10^{-7}\) is generally accepted as genome-wide significance of association and values between \(1 \times 10^{-5}\) and \(1 \times 10^{-7}\) as suggestive.

GWAS have demonstrated to be very sensitive to the introduction of bias generated by missing genotype data, genotype errors, cryptic relatedness between study subjects and population stratification (PS), which might cause spurious associations (67). PS refers to the systematic differences in allelic frequencies between cases and controls due to ancestry differences, which can give rise to false signals of disease association (68). Several methods have been developed to overcome PS. In GWAS, the association statistics can be adjusted by the genomic control (\(\lambda_{\text{GC}}\)), which measures the amount of inflation due to confounders such as PS. Alternatively ancestry can be inferred and accounted for by methods such as structured association and principal components analysis (PCA) (68). The transmission disequilibrium test (TDT) is an association test based on families instead of cases and controls, avoiding the problem of incorrect matching. TDT treats the allele that is transmitted to (one or more) affected children from each parent as a ‘case’ and the untransmitted alleles as internal ‘controls’ avoiding the effects of population stratification (69). Nevertheless, reported associations should be always taken with caution until multiple replications in independent sets of cases and controls and/or by independent researchers have been carried out.

The most important results from GWAS have been the elucidation of the cell types and the functional pathways that are involved in disease pathogenesis. Some associated genes were previously known by candidate gene studies, but most new associations did not involve obvious functional candidates.
A model explaining the role of the susceptibility loci in the pathogenesis of SLE has been proposed (19). A first group include variants leading to impaired immune clearance of apoptotic particles and nucleic-acid–containing immune complexes (DNASE1, TREX1, FCGR genes, C1, C2, C4, ATG5), which may induce the enhanced activation of plasmacytoid dendritic cells (pDCs) and autoreactive B cells, leading to the excessive production of type I IFNs and the expansion of autoreactive effector cells, respectively. A second group involves several susceptibility genes that affect the production of (TLR7, TLR9, IRF5, IRF7, SPP1 IRAK1, TNFAIP3, TNIP1, IFIH1) and the response to type I IFNs (TYK2, STAT4, MIR146A) (70). The abnormal function of innate immune cells may in turn activate the adaptive immune cells and both systems contribute to the inflammatory response and tissue damage. A third group of polymorphic genes are involved in ligand recognition, receptor signalling, activation of transcription and other immunological functions of B and T cells (HLA genes, BANK1, BLK, LYN, RASGRP3, AFF1, IKZF1, PTPN22, IL10, PCDCD1, FCGR2B, STAT4, TNFSF4, ETS1, BAFF) and may act by modulating the activation threshold, differentiation, expansion, cytokine production and other responses of these cells. The aberrant activation of the adaptive immune system results in loss of tolerance and the production of autoantibodies, which bind to nuclear antigens and further activate innate immune cells, creating a vicious circle that continuously amplifies the pathogenic processes in SLE. Additional risk variants may contribute by influencing the production of pro-inflammatory and other cytokines that mediate the tissue damage (TNF, IL17, IL18, KLK genes, FCGR3A, ITGAM, ACE).

Variation in the MHC genes represents by far the strongest and most consistent risk factor for SLE. However, the unusual long range LD of this genomic region that spans ~7.6 Mb and contains around 400 genes, makes it difficult to achieve the resolution needed to identify the primary loci responsible the disease susceptibility. Within the MHC, the strongest contribution to risk is given by the Class II genes HLA-DRB1 and HLA-DQB1, which code for the hyper-variable region of the HLA-DR and HLA-DR molecules β chain, respectively (71). The HLA haplotypes associated with SLE in Caucasians are DRB1*1501(DR2)-DQB1*0602, DRB1*0301(DR3)-DQB1*0201, and DRB1*0801(DR8)-DQB1*0402 (71). The association of other HLA-DRB1 alleles are more population-specific (72). Several other effects within the HLA have been reported (TNFA, C1, C4, C2), and dense mapping of the MHC in extensive collections of cases and controls have provided strong evidence of independent signals within the class I (RNF39-TRIM31), class III (SKIVL2, NOTCH4), and class II regions (HLA-DQB1-DQA2) (73, 74).
The role of many of the associated genes in SLE pathogenesis remains unknown (Table 2). Moreover, even for the genes that have been attributed to pathogenic pathways, the functional effects of most of the risk variants remain to be elucidated. Importantly, GWAS are based on tag SNPs, consequently, the associated SNPs may have a causal role (direct association) or, the associated SNPs may not actually be causal, but are in LD with the causal variant or variants (indirect association). Thus, it is possible that the association of genes with unknown functions represents the association of neighbour genes in LD. For example, the association of \textit{PHRF1} (\textit{KIAA1542}) is believed to be due to LD with \textit{IRF7} (55, 75).

**The pursuit of functional variants and the “missing heritability”**

The ultimate aim of genetic mapping is to understand the relationship between genotype and phenotype. Therefore, a list of associated variants without functional links has limited value. It is only through the identification of functional variants and/or the characterization of the effect of the risk alleles or haplotypes, that we can gain an understanding of the molecular pathways underlying the disease, and maybe manipulate them. Consequently, further fine mapping, haplotype analysis, or re-sequencing complemented with functional studies should follow the identification of all susceptibility loci. That has been the challenge and the motivation for this doctoral thesis.

**Selection of functional candidates**

To select the best candidates to be functionally tested among the many associated variants in LD is not a straightforward task. From a statistical point of view, conditional association analysis can identify distinct SNPs that have an independent effect, whose association does not depend on the co-variability (LD) between them. In addition, several conditional haplotype analysis strategies (76) can e.g. test whether the association of a particular haplotype background remains significant or not after changing the allele(s) only at a particular SNP or set of SNPs. However, even after careful haplotype analysis, it is often nearly impossible to distinguish the true functional variant among SNPs that are in very high LD ($r^2=0.8$) simply by means of statistical methods. The functional annotation of variants can aid in the selection process. Coding variants are the most obvious candidates as they affect the gene protein product. Non-synonymous (missense) changes can be prioritized by means of algorithms that give a score based on the potential deleterious effect of the amino acid change on the protein encoded. It is recommended to combine the predictions from multiple algorithms (SIFT, PolyPhen2, MutationTaster, LRT) instead of relying on a single one, along with a conservation score (PhyloP, GERP) (77). Variants located at exon-intron junctions and promoter sequences are also good candidates as
they may alter the splicing efficiency of the transcript and the level of gene expression, respectively.

Most SLE-associated variants are located in non-protein coding regions. The ENCYclopedia Of DNA Elements (ENCODE) Project is dramatically changing the way we look at the genome (78). The results from the ENCODE project suggest that the non-coding regions of the human genome are enriched for functional elements, including non-coding RNAs (62% of the genome), specific histone changes associated with promoters, enhancers, or other regulatory elements (56.1%), transcription factor binding sites (TFBSs, 8.1%), open chromatin sites (DNAse I hypersensitivity sites, 15.2%), and methylation of CpG sites (78). These numbers exceed by several folds the mere 2.94% of the genome that encodes protein-coding exons. Furthermore, according to the ENCODE data, 71% of the disease-associated loci reported in the GWAS have a potential causatively associated SNP in a DNase I site, and 31% of loci have a candidate associated SNP that is annotated to a TFBS (78). Thus, many of the non-coding variants identified by GWAS and/or their LD proxies are likely to be regulatory variants. The vast amount of data generated by ENCODE provides a tremendous resource for future association studies as a tool to discriminate potential functional variants and their functional effects.

**Imputation**

A major breakthrough in genetics has been the development of massive parallel sequencing or next-generation sequencing (NGS) technologies, which has made it possible to sequence thousands of genomes of individuals from different populations and to extensively catalogue the variation present in human populations (79, 80). Together, the 1000 Genomes Project (80) and the NHLBI 6500 Exomes Sequence Project (ESP) (81) have identified nearly 40 million single nucleotide variants (SNVs). Unlike common SNPs, low frequency SNPs and rare SNVs are much less shared across populations, highlighting the importance of including samples that represent diverse ancestries (49). The availability of such robust reference panels containing thousands of individuals densely genotyped has made possible the imputation, or prediction of non-typed genotypes, in a study sample. Imputation methods work by identifying shared haplotypes between the study individuals and the haplotypes in the reference panel and uses this information to “fill up” the most likely genotypes in the study individuals (82). Imputed genotypes can be then tested for association by adjusting for their inherent uncertainty. By this means, the coverage of a fine mapping can be significantly increased as well the statistical power and the probability to find the true causal variants. Imputation also makes possible the combination of genotype data from different studies generated by different platforms.
The genetic variants identified so far by GWAS of SLE and other common diseases explain only a small fraction (10-15%) of the familial aggregation (55, 75), which has incited a large discussion of the whereabouts of the “missing heritability”. The heritability is the proportion of the overall variance of a character that is due to genetic differences and the estimates of the variability explained by common risk variants are based on their additive genetic effects. So, a larger proportion of the heritability might be explained by gene-gene interactions, gene-environment interactions or epigenetic factors (83).

**Epistasis**

It has been suggested that epistatic associations, which have not been accounted for in the heritability estimations, could be responsible for a fraction of the unexplained heritability (84, 85). Genes can interact either in an additive or epistatic fashion. Epistasis in the statistical genetic sense takes place when the effect of interacting genes on a given phenotype differs from the expected sum of the effect of each gene taken individually (termed "additive" effect or "main" effect) (86). Synergistic epistasis occurs when the effect size of the interaction is greater than the expected additive effect, as opposed to antagonistic epistasis where the effect of the interaction is smaller. The term functional epistasis is also used; it addresses the molecular interaction that proteins have with one another, whether these interactions consist of proteins that operate within the same pathway or of proteins in complex with one another.

Several statistical methods to test for epistatic interactions have been developed (87). However genome-wide epistatic association studies remain a challenge, especially because of the enormously high amount of required tests ($2 \times 10^{12}$ tests for 2 million genetic variants), which further increases the burden of multiple testing on spurious associations. The use of prior biological knowledge can reduce the number of tests to be performed. For example, by limiting the test only to genes with prior evidence of functional interactions (88). On one hand, this approach strengthens the confidence of the findings and provides a more easily testable biological model, while on the other hand it lacks the ability to reveal novel interactions and pathways. A comprehensive study found that the patterns of shared susceptibility loci across ADs cluster the risk variants into groups and the proteins encoded near the risk variants in the same group tend to interact (52). Thus, genetic variation may influence several members of entire pathways increasing the risk to multiple diseases.

**Common disease-common variant hypothesis vs. rare allele model**

The reasoning behind GWAS was largely based on the common disease-common variant hypothesis (CD-CV), which proposed that the susceptibility
to common diseases is the result of ancient and therefore common variants (MAF >1%), which have escaped purifying selection because the risk alleles are only mildly deleterious or because of balancing selection (89). Supporters of the CD-CV hypothesis have the success of GWAS in the identification of multiple reproducible associations. The hope is that as more variants in the 1-10% range are catalogued and incorporated to GWAS they will be able to detect even more associated loci explaining a greater proportion of the heritability. Since this frequency class is enriched for variants under mildly deleterious selection, it probably contains most of the disease-associated alleles (3). The question remains about how feasible is to collect the amount of cases and controls necessary to detect those effects.

On the other side, the rare allele model proposes that the genetic variants underlying complex diseases are of recent origin and although individually rare (MAF <1%), as a group they are not. Mutations have various degrees of deleterious effects; at one end of the spectrum single lethal mutations would be enough to cause Mendelian diseases, while at the other end multiple mildly deleterious variants cause polygenic diseases. A few rare deleterious variants may be important and sufficient for an individual, or for a family, but not for all patients with the disease and, therefore, will be missed by GWAS (83). Very mildly deleterious variants can persist for a long time, enough to cause synthetic associations with common haplotypes detected by GWAS (90). Intermediate deleterious variants are eventually removed but the removal is balanced by de novo mutations. This hypothesis implies high allelic and genetic heterogeneity and requires a completely different study approach. The traditional case-control design may be inappropriate, as thousands of samples would be required to detect the effect of a rare variant. Whole-genome, exome, or targeted sequencing of associated genes is necessary. Careful selection of cases to be sequenced (i.e. those severely affected, carriers of the risk haplotypes, affected relatives) as well as association methods that consider the overall gene load of rare deleterious variants instead of their individual allelic frequency are required. Using this approach, rare and highly penetrant mutations associated with SLE have been identified in TREX1 (91), DNASE1 (92) and DNASE1L3 (93) genes.

Lastly, it is important to remember that SNPs have been the genetic markers of choice mainly because of their practicality: they are stable, abundant, their genotyping is adaptable to high-throughput systems and usually have a short-range LD, which limits the association signal to a ~100 kb region. But SNPs are not the only source of genetic variation and increasing evidence supports that structural variants such as CNVs are also important susceptibility factors for common diseases. CNVs of FCGR3B are associated with SLE (94). As many CNV overlap genes, they may have functional effects due to variable gene dosage.
Non-genetic factors

As SLE is a complex disease, the genetic factors only explain the disease to certain extent and it is clear that non-genetic factors also contribute. The ultraviolet (UV) radiation is the environmental factor most frequently associated with SLE (17). Exposure to UV light triggers disease flares and, in Sweden, a history of sunburn-susceptible skin type is associated with SLE risk (95). Also, as 90% of the patients are women and the disease activity increases during pregnancy, a role for female hormones has been proposed. Lower levels of androgens, higher levels of oestrogen and hyperprolactinemia have been described in cases with SLE (95). However, the sex bias may also be due to a gene dose effect of the X chromosome as supported by the increased prevalence of SLE among men with Klinefelter syndrome (46,XXY) (96-111). Viral infections are often associated with the disease onset and flares, in particular Epstein-Barr virus (EBV) infection in paediatric patients (112). Finally, drugs such as hydralazine, procainamide, isoniazid and minocycline can cause a type of lupus named drug-induced lupus erythematosus (DILE), which is characterized by predominance of skin manifestations and usually resolves after the drug has been discontinued (113).

The mechanisms by which non-genetic factors influence the development of SLE are even less understood. They may do so by inducing epigenetic changes in the genome. For example, hydralazine and procainamide (114) as well as UVB radiation inhibits DNA methylation in cells of patients with SLE (115). A study in 807 monozygotic discordant twins for SLE reported different methylation patterns in ~50 genes involved in immunological functions and folic acid metabolism, as well as a global pattern of hypomethylation in PBMCs (116). Future large-scale epigenetics studies may contribute to the understanding of SLE pathogenesis and the use of epigenetic modifying drugs has already been suggested as a novel therapeutic strategy in SLE (117).
The present investigation

Aims

1. The general aim of the present investigation was to characterize the association of susceptibility genes for SLE with focus on the identification of functional variants.

2. The papers included in this thesis were specifically aimed to validate, augment and refine the association of SNPs in the TNFSF4, STAT4, CD226 and BLK genes with the susceptibility to develop SLE, and to identify the causal variants underlying the associations.

Methods for Genetic Mapping of Disease Variants

Study subjects

This thesis comprises cross-sectional case-control studies in which genomic DNA from patients with SLE and ethnic-, gender- and age-matched controls were drawn from a collection of patients named BIOLUPUS. The collection was created and is coordinated by Prof. Marta Alarcón Riquelme at Uppsala University in Uppsala, Sweden. Sets of patients and controls were included depending on the availability of DNA samples at the time of each study. BIOLUPUS comprises around 1500 patients and 1500 controls of European ancestry collected from eight different centres in Europe and Argentina (Table 3) (118, 119). Paper II includes two replication sets from Mexico. Paper IV includes an additional collection of 279 patients and 515 controls from Scandinavia (44) and a European-American multi-centre collection. All patients fulfilled at least four of the 1982 American College of Rheumatology criteria for classification of SLE (15). All individuals provided informed consent as approved by the recruiting site institutional review boards at each of the participating institutions.
Table 3. Sample size and country of origin of the studied populations

<table>
<thead>
<tr>
<th>POPULATION</th>
<th>PAPER I (TNFSF4)</th>
<th>PAPER II (STAT4)</th>
<th>PAPER III (CD226)</th>
<th>PAPER IV (BLK)</th>
<th>PAPER V (BLKxBANK1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cases</td>
<td>Controls</td>
<td>Cases</td>
<td>Controls</td>
<td>Cases</td>
</tr>
<tr>
<td>BIOLUPUS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Argentina</td>
<td>278</td>
<td>355</td>
<td>171</td>
<td>171</td>
<td>60</td>
</tr>
<tr>
<td>Germany</td>
<td>259</td>
<td>322</td>
<td>247</td>
<td>220</td>
<td>284</td>
</tr>
<tr>
<td>Italy</td>
<td>258</td>
<td>223</td>
<td>221</td>
<td>207</td>
<td>264</td>
</tr>
<tr>
<td>Spain</td>
<td>517</td>
<td>901</td>
<td>390</td>
<td>620</td>
<td>206</td>
</tr>
<tr>
<td>Belgium</td>
<td>70</td>
<td>59</td>
<td>70</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>Hungary</td>
<td>38</td>
<td>35</td>
<td>38</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>Sweden</td>
<td>78</td>
<td>391</td>
<td>78</td>
<td>391</td>
<td></td>
</tr>
<tr>
<td>Portugal</td>
<td>163</td>
<td>165</td>
<td>163</td>
<td>165</td>
<td></td>
</tr>
<tr>
<td>Subtotal</td>
<td>1312</td>
<td>1801</td>
<td>1029</td>
<td>1218</td>
<td>1163</td>
</tr>
<tr>
<td>Mexico adults</td>
<td>231</td>
<td>250</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mexico paediatric</td>
<td>321</td>
<td>383</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scandinavian</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>USA</td>
<td>279</td>
<td>515</td>
<td>621</td>
<td>774</td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td>1312</td>
<td>1801</td>
<td>1581</td>
<td>1851</td>
<td>1163</td>
</tr>
</tbody>
</table>

The table shows the final number of cases and controls included in each study after data quality control.

Selection of SNPs to be genotyped

In an attempt to efficiently capture most of the common variation present in the genes of interest, a tag-SNP strategy was applied. Genotype data for the gene loci and the flanking upstream and downstream regions were downloaded from the HapMap CEU population (Utah residents with ancestry from northern and western Europe, data release 27, http://hapmap.ncbi.nlm.nih.gov/) (45). The HapMap genotype data was analysed in Haploview (120) to select tag SNPs and evaluate their performance. Any SNP with a minor allele frequency (MAF) below 5%, or showing Mendelian inconsistencies in the CEU trios and/or deviating from Hardy-Weinberg equilibrium (HWE) was excluded. Haploview tagger was run on aggressive mode (use 2- and 3- marker haplotypes) to increase tagging efficiency (121), and all alleles to be captured were correlated at an r^2>0.8 at least with one tag in the set. SNPs previously reported as associated with SLE or other ADs were included into the list of tags as well as SNPs with potential functional effects, i.e. non-synonymous mutations, location on regulatory elements, conserved regions, splicing and TFBSs.

Genotyping

All the SNPs on TNFSF4 (Paper I) and STAT4 (Paper II) were genotyped by using TaqMan® allelic discrimination assays in a 7900HT Fast Real-Time System (ABI, Foster City, CA). SNPs on CD226 (Paper III) and BLK (Paper
IV and V) were genotyped at an external facility by using a GoldenGate® Custom Genotyping Assay and a BeadXpress Reader (Illumina, San Diego, CA). In some cases, genotyping of additional SNPs and/or samples was performed using TaqMan®. Genotype data for *BANK1* and *BLK* (Paper V) was extracted from a previous genome-wide association study in which samples were genotyped using the Affymetrix GeneChip® Human Mapping 100K set (Affymetrix, Santa Clara, CA).

TaqMan® assays (*Figure 4*) rely on the 5´–3´ exonuclease activity of the Taq polymerase and the principle of Fluorescence Resonance Energy Transfer (FRET). A SNP assay contains a pair of primers that amplify the target sequence where the SNP is located and two TaqMan® allelic-specific oligonucleotides (ASO). Each ASO has a distinctive fluorophore (donor of fluorescence when excited) attached to the 5’-end and a common quencher (absorbs the fluorescence by FRET) at the 3’-end. FRET takes place only when the fluorophore and the quencher are in close physical proximity as it occurs when the probe is intact. During DNA amplification, the ASOs hybridize with their corresponding allelic-specific complementary sequence. As the Taq polymerase extends the primers, it also cleaves the allele-specific 5’ fluorophore, which stops the FRET and release a fluorescent signal. After the PCR is completed, genotype for each SNP is assigned by measuring the allele-specific fluorescence.

*Figure 4. Allelic discrimination achieved by the selective annealing of TaqMan® probes.* Copyright © 2012 Life Technologies. Reproduced with permission.
The GoldenGate® system combines allelic-specific primer extension and DNA ligation for allelic discrimination. For each SNP, the reaction contains two ASOs (P1 and P2) and one locus-specific oligonucleotide (LSO) (P3). Each ASO contains a PCR primer site that hybridizes to one of two universal dye-labelled primers. The LSO also contains a universal PCR primer site and a unique tag sequence. The allele-specific primer-extension of the correctly matched ASO followed by ligation to the LSO, creates a substrate for PCR using universal PCR primers P1, P2, and P3. The single-stranded, dye-labelled products are then hybridized through their unique tag sequences to a specific bead in a BeadArray. After hybridization, a BeadArray Reader is used to analyse fluorescence and call genotypes (122).

In Affymetrix’s GeneChip® Mapping Assay, genomic DNA is digested with a restriction enzyme (XbaI or HindIII) and then the fragments are ligated to universal adaptors recognizing the overhangs. A universal pair of primers recognizing the adaptor sequence is used to amplify the adaptor-ligated DNA fragments. The PCR products are then fragmented, labelled and hybridized to oligonucleotides immobilized on the Mapping 100K Set microarray. The immobilized oligonucleotides are designed to distinguish alleles for specific SNPs (122).

Statistical analysis

Genetic epidemiology and statistical genetics have adapted several statistical methods from traditional epidemiology to understand the relationship between phenotype and genotype. While traditional epidemiologists study the effect of certain exposures (i.e. tobacco) on the risk to develop a disease, genetic epidemiologist study the effect of carrying certain genetic variant(s).

In brief, association studies compare the allelic, genotype and haplotype frequencies at selected polymorphisms between patients and healthy controls. The specific statistical methods used for each paper are mentioned under the corresponding materials and methods section and an overview of the methods applied is given below.

Single marker analysis

For each single diallelic SNP with alleles A and a, and assuming a as the risk allele, differences in allele frequencies can be determined by 2x2 contingency tables using the traditional \( \chi^2 \) or Fisher’s exact tests (df=1) and/or logistic regression methods. However, the results of allelic tests need to be taken with caution because in reality alleles at one polymorphic site are not independent of each other. It is the combination of both alleles that together exert the effect on the liability to develop the disease, and they may be acting in an additive, dominant, co-dominant or recessive fashion. In
addition, for many genetic statisticians, allelic tests overestimate the association because they duplicate the sample size (N individuals = 2N alleles). Consequently, genotypic tests give a better estimate of the effect of a variant. In this thesis, the Armitage’s trend test and logistic regression genotypic tests were widely used. The Armitage’s trend test considers the genotypes as several levels of exposure to the risk allele (0= AA, 1= Aa and 2= aa) and assumes an additive genetic model, that is, an ascending “trend” of the risk as the number of risk or a alleles increases (123). When we interrogated the different models of inheritance for particular markers, we derived contingency tables for each model (i.e. dominant, recessive), analysed them using either $\chi^2$ or logistic regression and selected the best-fit model (123).

**Haplotype analysis**

The tests used for haplotype analysis are similar to the ones use for single-marker tests. As the data consisted of unphased genotypes, the gametic phase of the alleles had to be estimated first. The genetic software used (Haploview and PLINK) use the Expectation Maximization (EM) algorithm, an iterative process that obtains maximum-likelihood estimates of haplotype frequencies from multi-locus genotypes (124).

Haplotype and genotype association tests are complementary. The haplotype analysis reveals critical information about the LD structure of the gene and the story of the associated variants. It is especially useful for studies based on tag SNPs because in most of the cases they are not causal variants, as suggested by their name. When a new mutation arises, the association between each mutant allele and its ancestral haplotype is disrupted only by mutation and recombination in subsequent generations. Thus, it is possible to track risk alleles in the population by identifying (through the use of tag SNPs) the particular haplotype block or gene segment on which they arose (125). To narrow down or refine the associated gene segment, sliding windows of two, three or more SNPs were tested (Papers III and IV). PLINK offers a wide and flexible menu of conditional haplotype tests that make possible to test whether specific haplotype(s) or SNP(s) can explain the association of a haplotype block (76).

**Linkage Disequilibrium**

LD refers to the statistical association between alleles at two or more variable sites (i.e. SNPs) along the same chromosome at the population level. If we consider two diallelic SNPs A>a and B>b, the frequencies of the four possible haplotypes or allele combinations (AB, Ab, aB and ab) are expected to become the product of the individual allele frequencies after many generations of recombination between the two sites. Any departure from this is called linkage disequilibrium and is defined by $D = (AB)(ab) – (Ab)(aB)$. The statistic $D'$ is the absolute value of $D$ divided by the
maximum value that $D$ could take given the allele frequencies and ranges between 0 (no LD) and 1 (complete LD). Another way of measuring LD is by the coefficient of determination between alleles ($r$). The values of $r^2$ (the square of the correlation coefficient) also lie between 0 and 1 but differs from $D'$ in that its maximum possible value depends on the minor allele frequency of the two SNPs. The advantage of $r^2$ over $D'$ is that it can measure how well one SNP can act as a surrogate (or proxy) for other SNP.

**Epistasis**

In Paper V we used the method developed by Wirapati et al. (126) to test for synergistic epistasis between BANK1 and BLK. The method is summarized in Figure 5. If two genotypes, when combined, had a significant association ($S$ score significant, $P<1\times10^{-5}$) but there was no significant epistatic effect ($P_e>1\times10^{-5}$), we concluded that such an association was mainly due to the sum of the individual or marginal effects of the associated genotypes. If the epistatic effect was significant ($P_e<1\times10^{-5}$), we then refer to it as a genetic epistatic interaction.
Figure 5. **Method to detect epistasis developed by Wirapati et al.** (126) A) For each pair of non-correlated SNPs \(r^2<0.8\), 16 contingency tables formed by the possible combinations of genotypes at both SNPs under dominant and recessive models are recorded. The figure shows one of 16 possible combinations. B) For each contingency table, a Pearson score \(S\) is computed with its corresponding \(P\) value. Here, a significant association reflects the sum of additive (or main) and epistatic effects for a specific genotype combination (dominant or recessive). To determine the epistatic effect, that is, an association odds ratio higher than expected under the null hypothesis of independence, an epistatic score is derived from the difference between the observed Pearson score \(S\) of each contingency table and the expected Pearson score \(S_0\) under the null hypothesis of no epistasis. An epistasis \(P\) value \((P_e)\) is obtained through permutation.
Quality control of the data

Association tests are very sensitive to the bias introduced by confounding factors. In order to minimize bias several strategies were applied. Duplicated and/or related samples were detected by IBS analysis and removed, as well as samples with individual missing genotyping call rates higher than 10%. Only SNPs with genotyping call rates over 90% that follow HWE in controls were included in the analysis. Only alleles and haplotypes with allelic frequencies over 5% were analysed, except in Paper IV (>0.5%). Multiple testing was corrected by False Discovery Control or permutation analysis.

Patients and controls coming from different countries were analysed separately to avoid spurious associations due to population stratification. A meta-analysis was used to summarize the association by introducing the country of origin as a stratification variable. In a meta-analysis, the association is calculated within each stratum first and then an average OR weighted by the sample size of each stratum is estimated as well as a 95% confidence interval. This method is known as a Cochran-Mantel-Haenszel meta-analysis (25, 29) and it assumes that the allelic effects are the same in all the studies (countries in these investigations). Therefore, it is also known as fixed-effect meta-analysis. A test of the homogeneity of the OR across strata was also included (Breslow-Day test) (120). In the presence of a significant P-value indicating heterogeneity of the association, a random-effects meta-analysis was applied by the DerSimonian-Laird test. The DerSimonian-Laird test incorporates in the meta-analysis the assumption that the studies are not all estimating the same allelic effect (127).

For the individuals included in the studies described in Papers III, IV and V, genotype data from 350 ancestry informative markers (AIMs) was available. AIMs are genetic markers that display large differences in allele frequencies between known human subpopulations and are used to infer genetic ancestry (68). We estimated the contribution of European, Asian, and African ancestry of our study subjects by a model-based clustering program (STRUCTURE)(128) using the HapMap Phase I populations as reference. Only individuals having more than 90% of European Ancestry were retained.
Results and Discussion

Paper I: Replication of the TNFSF4 (OX40L) Association

The OX40 ligand (OX40L) is a membrane-bound protein expressed on the surface of antigen presenting cells (APCs), including activated B cells and DCs, while its cognate receptor is primarily expressed on activated T cells (129). The interaction OX40-OX40L represents a co-stimulatory signal to activated T cells at effector and memory stages of the immune response, which is critical for their survival and clonal expansion (130). The reverse signalling via OX40L, in turn, enhances B cell proliferation and differentiation (131). OX40L also inhibits the generation and suppressive effects of type 1 regulatory T cells (Tr1) (132).

The TNFSF4 gene (Tumour Necrosis Factor (ligand) Superfamily, member 4) codes for OX40L and is located in a SLE linkage region on 1q25 (133, 134). TNFSF4 was selected as the most plausible candidate gene underlying the linkage signal based on its function (27). Thus, it cannot be discarded that other genes within 1q25 may be contributing to the linkage observed in SLE. The first association study of TNFSF4 for SLE reported strong evidence of association with a 80kb haplotype in the 5’ region of the gene in UK and US Minnesota populations (27). Graham et al (27) observed a risk haplotype tagged by the minor alleles of SNPs rs1234317-T, rs2205960-T, rs12039904-T, or rs10912580-G, and what seemed to be a protective haplotype tagged by the minor allele of rs844654-A.

The present investigation aimed to validate the association of TNFSF4 and SLE as no replication studies had been published. Since there was no evidence supporting any of the associated SNPs as the single causal variant, we selected tag SNPs capturing all the common alleles in the 80kb associated haplotype block from the HapMap-CEU population LD structure. In total, five SNPs were tested for association including three SNPs from the previous study (rs1234317, rs844654 and rs12039904). We did not include SNPs rs2205960 nor rs10912580 because rs12039904 was an excellent surrogate for them with $r^2=0.90$ and $r^2=0.95$, respectively. The results are presented in Paper I tables 1 and 2, and can be summarized by Figure 6.
Figure 6. Haplotype association of TNFSF4. We observed a unique risk haplotype bearing the minor alleles of rs1234317-T and rs12039904-T corresponding to the same haplotype reported by Graham et al (27). This haplotype was consistently associated through the different populations (P=8.00x10^{-5} Argentina, P=1.86x10^{-2} Germany, P=5.80x10^{-6} Italy). However, we found different “protective” haplotypes depending on the population. As shown in the figure, all these protective haplotypes share the major alleles rs1234317-C and rs12039904-C.

These results suggested that rs1234317 and rs12039904 explained the association of the risk and protective haplotypes. In order to confirm this, we used a conditional haplotype test implemented in PLINK v1.03. In brief, to determine whether one or more SNPs explain the association of a haplotype, it is possible to perform a general or “omnibus test” for the haplotype before and after controlling for the SNP(s) of interest (Table 4).

Table 4. Conditional haplotype analysis of TNFSF4 controlling for rs1234317 and rs12039904.

<table>
<thead>
<tr>
<th>Population</th>
<th>Haplotype omnibus test P-value</th>
<th>Haplotype omnibus test P-value after controlling for rs1234317 and rs12039904</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>8.00x10^{-5}</td>
<td>0.3670</td>
</tr>
<tr>
<td>Germany</td>
<td>1.86x10^{-2}</td>
<td>0.0953</td>
</tr>
<tr>
<td>Italy</td>
<td>5.80x10^{-6}</td>
<td>1.0000</td>
</tr>
<tr>
<td>Spain</td>
<td>0.1129</td>
<td>NA*</td>
</tr>
</tbody>
</table>

If the omnibus test were still significant after controlling for SNPs rs1234317 and rs12039904, then other SNPs would still be contributing to the association. However, there was no association in any population after the adjustment and thus rs1234317 and rs12039904 explained the whole
haplotype effect. The test was not performed in Spain because conditional tests are only justified when the omnibus is significant.

*TNFSF4* is today one of the shared susceptibility genes for ADs. Its association with SLE has been confirmed by several studies (28, 57, 135, 136), and it seems to play a role in MS (137), primary Sjögren’s syndrome (pSS) (138), Crohn’s disease (ChD) (139), and Celiac disease (CeD) (140). Although different SNPs have been associated with SLE, their risk alleles tag the same risk haplotype that we reported in our study, in other words, they are in complete LD. The original association study by Graham *et al* (27) reported that on EBV-activated lymphoblastoid cell lines (LCLs) and peripheral blood lymphocytes (PBLs) homozygous for the risk haplotype the expression of OX40L on the cell surface and of *TNFSF4* mRNA were significantly increased. Although a functional effect of the haplotype has been demonstrated, no causal variant has been identified. However, a recent study demonstrated that the risk allele of rs2205960-T creates a binding site for the NFκB RELA-p50 heterodimer resulting in higher NFκB in *vivo* binding (141). This SNP is located in a DNase hypersensitivity site and is the most likely candidate causal variant.

The mechanism by which the overexpression of OX40L and the secondary enhancement of OX40-OX40L signalling contribute to the disease pathogenesis remains to be elucidated (129). The OX40L overexpression on B cells may contribute to the aberrant activation and proliferation of B cells characteristic of SLE. In OX40⁺ CD4⁺ T cells, homozygosity for the risk haplotype may enhance APC–T cell interactions resulting in enhanced co-stimulation. Alternatively, the overexpression of OX40L may hinder the generation of IL-10–producing Tr1 cells, which are critical inhibitors of autoimmunity (129).

The number OX40⁺ T cells correlates with disease activity, constituting it as a potential SLE biomarker (142). An increased number of OX40⁺ T cells and increased serum levels of OX40L are observed in patients with SLE when compared to controls, particularly in patients with lupus nephritis (142). Interestingly, association studies that use the status of different ACR criteria as sub-phenotypes have found an association between *TNFSF4* and renal disorder (143, 144). The potential role of OX40L in the pathogenesis of lupus nephritis is supported by the abundance of OX40L observed along the epithelial side of the glomerular capillary wall in patients with proliferative lupus nephritis, where it co-localizes with immune complexes (145). The pathological mechanisms through which the signalling OX40L-OX40 contributes to renal damage remain to be clarified but they represent a promising target for therapy (146).
Paper II: Two Independent Effects in STAT4 that Correlate with Gene Expression and Act Additively with IRF5

**STAT4** (Signal Transducer and Activation of Transcription 4) is a transcription factor important for the transduction of signals induced by type I IFNs. STAT4 is expressed mainly in T cells, natural killer (NK) cells, activated monocytes and DCs (147), where it resides in the cytoplasm in its inactive form. In response to cell activation with IL-12, IL-23 or IFNα/β, STAT4 becomes phosphorylated by tyrosine kinases, forms homodimers or heterodimers with STAT1 and STAT3 and translocates into the nucleus, where it induces expression of IFNγ genes (148). STAT4 is essential for the differentiation of naïve CD4+ T cells into T helper type 1 (Th1) cells (149). Patients with SLE display abnormally high levels and activity of type I IFNs (150-153).

The **STAT4** gene (2q32) was first associated with ADs by Remmers et al (154) through the fine mapping of a RA linkage region in European-Americans. This study found strong association between RA and SNP _rs7574865_ located on the third intron of the gene, which was later associated with SLE (154). At the moment that the study by Remmers et al was published, we were following up the **STAT1-STAT4** region after having identified it in a 100,000 SNPs GWAS conducted in Argentinians. It was indeed exciting because until that moment the HLA was the strongest genetic factor reported for SLE in individuals of European ancestry, followed by **IRF5** and **ITGAM** (54, 55, 73, 119). However, in Argentinians, the association of the **STAT1-STAT4** region was stronger than the HLA. The associated SNPs were located on the 16th intron (62kb downstream from _rs7574865_) and on the intergenic region between **STAT1** and **STAT4** (Figure 7).

Paper II (155) aimed to validate the genetic association of **STAT4** in SLE and define whether the multiple associated variants reflected allelic heterogeneity or were due to LD. We conducted a fine mapping of the **STAT1-STAT4** region in an independent set of cases and controls from Spain. Four SNPs in **STAT4** remained associated after correction for multiple testing being _rs3821236_ the strongest associated variant (_P_=6.4 x 10^-6_), as in the Argentinian GWAS (155) (Paper II: Figure 1). The LD structure of the gene revealed that _rs3821236_ and _rs7574865_ were each representing the association of two separated although moderately correlated haplotype blocks, one covering from intron 8th to 16th and the other covering a segment on the 3rd intron, respectively (Figure 8) (Paper II: Figure 1 and Supplementary Figure 1).
Figure 7. The STAT1–STAT4 region in a GWAS of SLE. The figure shows the top associated SNPs in a GWAS carried out in 255 SLE cases and 256 controls from Argentina (191 cases and 141 controls after PCA analysis). The significance values (–log₁₀P-value) are corrected for stratification by using 149 AIMs extracted from the Affymetrix 100k array. The strongest signals of the whole array were observed with SNPs rs3821236 and rs3024886 located in STAT4, and SNP rs2030171 in STAT1. The complete STAT1–STAT4 region was followed up in Paper II (155).

Figure 8. Fine mapping of the STAT1–STAT4 region. The physical positions (top panel) of the SNPs typed in 390 patients and 480 controls from Spain covering the STAT1–STAT4 region are shown. Haplotype blocks 3, 4 and 6 (middle panel) were associated with disease susceptibility (³block 3 risk haplotype P-value = 0.0037; ⁴block 4 risk haplotype P-value = 2.64 × 10⁻⁶; ⁶block 6 risk haplotype P-value = 9.52 × 10⁻⁶). Risk haplotypes are shown in red and the top single-markers are underlined. The bottom panel shows the significance of the association data (–log₁₀ P-value) for 25 tag SNPs passing genotyping quality control. The blocks were defined using the solid spine of the LD method in Haploview version 4.0.
We then tested and compared the association of these four associated variants in populations of different origin, including Europeans (Italy, Germany, Spain, Argentina) and European-Amerindians (Mexico) (Table 5). All the SNPs were replicated in the meta-analysis, however, only rs3821236 and rs7574865 displayed genome-wide significance. While rs3821236 had the strongest effect in the Spanish and Argentinean sets, rs7574865 was the strongest variant in Italy and Mexico. A random effects meta-analysis was applied because some SNPs displayed heterogeneity of the OR.

Table 5. Analysis by country and meta-analysis of the four strongest susceptibility variants in STAT4 (Summary)

<table>
<thead>
<tr>
<th>Population/Test</th>
<th>rs1467199 (G)</th>
<th>rs3821236 (A)</th>
<th>rs3024866 (C)</th>
<th>rs7574865 (T)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Germany</td>
<td>0.940</td>
<td>0.99</td>
<td>0.305</td>
<td>1.18</td>
</tr>
<tr>
<td>Italy</td>
<td>0.842</td>
<td>1.03</td>
<td>1.73x10^{-05}</td>
<td>2.04</td>
</tr>
<tr>
<td>Spain</td>
<td>8.01x10^{-05}</td>
<td>1.50</td>
<td>2.71x10^{-09}</td>
<td>1.89</td>
</tr>
<tr>
<td>Argentina</td>
<td>0.085</td>
<td>1.33</td>
<td>6.53x10^{-05}</td>
<td>1.87</td>
</tr>
<tr>
<td>Mexico</td>
<td>0.122</td>
<td>1.33</td>
<td>1.46x10^{-05}</td>
<td>1.81</td>
</tr>
<tr>
<td>Random effects</td>
<td>1.41x10^{-02}</td>
<td>1.24</td>
<td>8.98x10^{-11}</td>
<td>1.75</td>
</tr>
<tr>
<td>meta-analysis</td>
<td>0.138</td>
<td>0.124</td>
<td>0.024</td>
<td>0.013</td>
</tr>
<tr>
<td>Heterogeneity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Then, we adjusted the association of all SNPs by the two strongest variants (rs3821236 and rs7574865) by using logistic regression analysis (Table 6). These results demonstrated that rs3821236 and rs7574865 represent two independent effects as their association remained significant after conditioning for each other, and that no other markers were independently associated with the disease.

Table 6. Conditional association analysis of STAT4 adjusting by the two more associated variants: rs7574865 and rs3821236

<table>
<thead>
<tr>
<th>Condition on</th>
<th>rs1467199 (G)</th>
<th>rs3821236 (A)</th>
<th>rs3024866 (C)</th>
<th>rs7574865 (T)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P-value</td>
<td>OR</td>
<td>P-value</td>
<td>OR</td>
</tr>
<tr>
<td>rs7574865</td>
<td>0.225</td>
<td>1.08</td>
<td>1.22x10^{-2}</td>
<td>1.21</td>
</tr>
<tr>
<td>rs3821236</td>
<td>0.365</td>
<td>1.06</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

Like STAT4, IRF5 was a confirmed susceptibility gene for SLE (54, 55, 119), and a signalling molecule involved in the cellular response to type I IFNs. Thus, we hypothesized that these genes might be interacting on the risk to develop SLE in an epistatic fashion. We observed an additive effect between STAT4 and IRF5 but not an epistatic interaction. By looking at the
area under the receiver operating characteristic (ROC) curve (C-statistic) we observed a maximum ability to discriminate between cases and controls for the combination of SNPs rs2070197 (IRF5) plus rs7574865 (STAT4) (Paper II: Figure 2).

To investigate the functional effect of STAT4 polymorphisms, mRNA expression levels were measured by quantitative PCR in peripheral blood mononuclear cells (PBMCs) from 73 healthy donors, who were genotyped for the associated variants. A general genotypic test performed by using ANOVA (ANalysis Of VAriance) revealed nominal association between the three SNPs in STAT4 and the mRNA levels ($P < 0.05$). The genotype at SNP rs3024866, which is closely linked to rs3821236, had the strongest effect on the mRNA expression level ($P=0.009$) (Paper II: Figure 3).

To date, STAT4 is a major autoimmune gene associated with several ADs in Caucasians (54, 154-160), Hispanics (156, 161) Asians (162-165), and African-Americans (63, 156). Importantly, differences in the LD structure of STAT4 across populations with diverse genetic ancestry have been described (156). A fine mapping of STAT4 validated the independent effects that we reported in Paper II in individuals of Hispanic and African-American origin (156). In Hispanics, compared to Caucasians, the LD between the two associated blocks was lower and the frequency of the exon 6-intron 16 haplotype was higher (155, 156). A GWAS in Behçet's disease in Asians (166) identified a disease associated block in the promoter of STAT4. The three SNPs forming this block showed no LD with the intron 3 SNP rs7574865, which was not found to be associated with the disease. Thus, the allelic heterogeneity of STAT4 likely reflects population differences, with different alleles being important for different populations. An alternative explanation for the observed allelic heterogeneity of STAT4 is that the different risk alleles reflect different clinical subsets of SLE patients. The association of the intron 3 variants with the presence of anti-dsDNA autoantibodies, nephritis, early disease onset and absence of oral ulcers supports this (63, 157, 162, 167). Unfortunately, in our cohort we did not have complete clinical data to test whether rs3821236 is associated with any particular SLE clinical subset. Furthermore, in RA patients, SNP rs7574865 is associated with both ACPA(autoantibodies to citrullinated proteins/peptides)-positive and ACPA-negative RA, with a stronger effect size in ACPA-negative RA (168). Whereas other SNPs in the STAT1-STAT4 locus are only associated with ACPA-negative RA. Thus, the allelic heterogeneity of STAT4 and possibly the intergenic region STAT1-STAT4 may be reflecting the clinical heterogeneity and further investigation could provide us with a better understanding of its role in the pathogenesis of ADs.
Genetics together with functional studies have improved our understanding of the mechanisms by which variation in \textit{STAT4} contributes to ADs. In Paper II we show that the haplotype risk correlates with higher levels of \textit{STAT4} expression (155, 157). Our results also suggest that the expression of \textit{STAT4} is regulated in a cell-specific manner as tissue-specific alternative promoters were identified (155). This complexity of gene regulation poses an additional obstacle in the search for the true functional variants. A recent study reported a correlation between rs7574865-T and increased expression of interferon-induced genes in the absence of high amounts of IFN (169). The SNPs reported in Behçet’s disease, which are located in a regulatory element, increase the expression of \textit{STAT4} and the production of IL-17 (166). Hence, \textit{STAT4} carries at least three risk loci for ADs, which together increase the sensitivity to IFNα (169) enhance \textit{STAT4} expression (155), and act in an additive manner with \textit{IRF5} to increase disease susceptibility, probably through the up-regulation of target genes like \textit{IL17}.

Both \textit{STAT4} and \textit{IRF5} are major hubs in the signalling pathways regulating the production and response to type I IFNs. By using ChiP-seq, Wang et al (170) demonstrated that \textit{IRF5} and \textit{STAT4} bind to 3,190 common targets in stimulated PBMCs, and their binding sites are located within 1kb from each other, which suggests a physical interaction. Several target genes of \textit{STAT4} and \textit{IRF5} belong to type I IFN-related pathways, of which many are associated with ADs (reviewed in (70)). They are believed to act together leading either to increased production of IFNs, amplification of the deleterious IFN-mediated signalling, impaired negative regulation of the IFN pathways, or to abnormal response to IFNs caused by the increased sensitivity of cells.
Paper III: A functional 3’UTR Variant in CD226 Impairs Gene Expression in T and NKT Cells

CD226 (also known as DNAX accessory molecule 1, DNAM1) is located in chromosome 18q22 and was first associated with ADs by a GWAS of non-synonymous SNPs in T1D (171). This study found strong evidence of association with the variant Gly307Ser (rs763361 G>A) located in the 7th and last exon of CD226. Further fine mapping of the gene in T1D and MS, suggested rs763361 as the causal variant underlying the association of CD226 (172). Hafler et al (172) used a forward multivariate logistic regression analysis in which the associated SNPs were added to the model already containing rs763361 under the assumption that if any SNP contributed significantly to the association of rs763361 alone, it then represented an independent effect. They acknowledged that although they found no evidence against the hypothesis that the non-synonymous SNP (Gly307Ser) is the causal variant, they could not exclude that another variant in LD with Gly307Ser was the true causal variant (172).

CD226 is a transmembrane protein of the Immunoglobulin (Ig)-superfamily containing two Ig-like domains. It is expressed on the surface of hematopoietic cells, where it mediates a wide range of immune functions including activation and differentiation in T cells, cytotoxicity activity in cytotoxic T lymphocytes (CTLs) and NK cells, natural killer T (NKT) cell apoptosis, platelet and megakaryocyte adhesion to endothelium, and monocyte extravasation (173, 174). CD226 also plays a role in the establishment of follicular T helper cells in the small intestine, maturation of CD8+ thymocytes, and inhibition of apoptosis in thymocytes (175, 176). The identified ligands of CD226 are the beta2-integrin LFA-1 (CD11a/CD18), the poliovirus receptor (PVR/CD155), and the PVR-related family-2 receptor (PRR-2/CD112/nectin-2), which are normally expressed on epithelial and primary vascular endothelial cells, and aberrantly expressed on many tumour cells. The interaction of CD226 with its ligands provides a co-stimulatory signal to T and NK cells and mediates the recognition and lysis of target cells (177).

At the time of the present study, no association of CD226 and SLE had been reported. CD226 was a strong gene candidate and growing evidence supported it as a shared autoimmunity gene. Thus, we performed a fine mapping of CD226 in European patients with SLE and controls (Paper III) (178). The LD structure of HapMap-CEU population showed that rs763361 was located on a long haplotype block covering the 3’ end of the gene and, therefore, we typed more SNPs in this region (Figure 9a).
Figure 9. Association of the 3-SNPs haplotype in the 3′-untranslated region of CD226 with SLE. a) LD structure of CD226, as determined using Haploview solid spine method. b) Analysis of haplotypes by sliding window analysis of 2-SNP haplotypes and 3-SNP haplotypes. The y-axis shows the $-\log_{10}P$ values only for the haplotype windows associated ($P < 0.05$) with SLE. The x-axis shows the 12 SNPs genotyped across the CD226 locus. Relative SNP positions are mapped on the gene, with coding exons shown in red and non-coding exons shown in blue. The arrow shows the direction of transcription. The enlargement of the 3′ region highlights the 3 SNPs giving the best association model: rs763361, rs34794968, and rs727088.
In total, twelve tag SNPs in CD226 were typed and tested for association by a Cochran-Mantel-Haenszel meta-analysis (Paper III: Table 1). The association of rs763361 was replicated in SLE ($P=1.74\times 10^{-03}$) but two other SNPs were also strongly associated with the disease, namely rs34794968 ($P=8.51\times 10^{-03}$) and rs727088 ($P=9.57\times 10^{-04}$). These three SNPs were in high LD and were part of a haplotype block covering the last exon of CD226 and ~5 kb downstream the gene (Figure 9a). To delimit the precise gene segment where the association signal was located, and to discern the causal variant of this block, we tested the association of each haplotype of two and three contiguous SNPs by a sliding window approach. This analysis pointed to the region between SNPs rs34794968 and rs727088, as both were included in the two best-associated windows (Figure 9b). The best model was given by the haplotype formed by SNPs rs763361-rs34794968-rs727088 (Omnibus test, $P = 8.63\times 10^{-4}$). These SNPs formed three common haplotypes: ATC, GGT and AGC (Table 7). By testing each haplotype against all others, only the ATC haplotype was reliably associated with SLE ($P=1.30\times 10^{-4}$, OR 1.24 (1.11-1.38)).

**Table 7. Association of the CD226 haplotype block rs763361-rs727088 covering exon 7 and the region downstream of the gene**

<table>
<thead>
<tr>
<th>rs763361</th>
<th>rs34794968</th>
<th>rs727088</th>
<th>Freq cases</th>
<th>Freq controls</th>
<th>P-value</th>
<th>OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GGT</td>
<td></td>
<td></td>
<td>0.479</td>
<td>0.511</td>
<td>0.100</td>
<td>0.91 (0.82-1.02)</td>
</tr>
<tr>
<td>ATC</td>
<td></td>
<td></td>
<td>0.443</td>
<td>0.397</td>
<td>1.30\times 10^{-4}</td>
<td>1.24 (1.11-1.38)</td>
</tr>
<tr>
<td>AGC</td>
<td></td>
<td></td>
<td>0.078</td>
<td>0.091</td>
<td>0.122</td>
<td>0.87 (0.70-1.00)</td>
</tr>
</tbody>
</table>

Due to the high LD between the three associated SNPs, it was not possible to test whether any of risk alleles had an effect independent of the haplotype background. We took the top associated SNP (rs727088) and added the genotypes of the other two SNPs as covariates in a forward step-wise multivariate logistic regression analysis and found that no other SNP contributed to the association. The question remains if this approach is really optimal for dissecting the causal variant among a set of highly correlated SNPs. We decided to study other endophenotypes, less complex than the disease status. By quantitative real-time PCR and FACS analysis, we analysed the mRNA and protein levels of CD226 in PBMCs from healthy donors, respectively. To individually evaluate the impact of each associated variant on gene expression, we used luciferase reporter constructs containing different portions of the CD226 3’UTR region. This way, we could study separately the effect of the associated variants in vitro. Homozygosity of the haplotype ATC correlated with lower expression of CD226 mRNA in PBMCs (Paper III: Figure 2a), as well as with reduced protein expression on the surface of CD4⁺, CD8⁺ T cells, and NKT cells, but not in NK cells.
(Paper III: Figure 2b). Finally, luciferase reporter assays with the deletion constructs indicated that only the allele rs727088-C resulted in a reduced expression of CD226 mRNA (Paper III: Figure 4a). This was not due to an accelerated degradation rate of the transcript (Paper III: Figure 4b).

In Paper III (178) we identified, for the first time, an association of CD226 with SLE in individuals of European ancestry. Further, our data strongly supports the 3’UTR SNP rs727088 as a candidate functional variant affecting the regulation of CD226 transcription specifically in T and NKT cells. The molecular mechanism by means rs727088 affects the transcription of CD226 remains to be elucidated and future studies are needed to address the demonstrated cell-specific effect of the haplotype. As we examined only SNPs, it remains a possibility that a different type of genetic polymorphism, in high LD with rs727088 and most likely within the 3’UTR, could be responsible for the lower transcript expression.

On the other hand, SNP rs763361 has been now widely associated with multiple autoimmune diseases including RA (172), MS (172), Wegener’s granulomatosis (WG) (179), CeD (180), AITD (172, 180) and SSc (181). In the latter case, the rs763361 association was confined to the diffuse cutaneous form (dcSS), which is considered to be more severe (181). In all these studies, only the SNP rs763361 was tested, and no comprehensive studies of CD226 have been performed except in T1D (171) and SLE (Paper III)(178). The SNP rs763361 is a good functional candidate because it is a non-synonymous variant. The risk T allele determines the presence of a serine instead of a glycine (C allele) in the amino acid position 307, which lies within the protein cytoplasmic domain and thus may modulate signal transmission and cell-activation (173). The T allele (Ser) is the ancestral allele in primates and the change to C (Gly) is predicted to be benign or tolerated by the algorithms implemented in PolyPhen2, SIFT, LRT, and MutationTaster (77, 182). It has also been hypothesized that this variant could disrupt an exonic splicing silencer (ESS), leading to alternative splicing of CD226. Although we found an alternative Δexon2-3 CD226 splicing isoform, it was expressed at very low levels (178). The Gly370Ser may be functional, yet it is not responsible for the decreased levels of expression of CD226 observed in individuals carrying the risk haplotype for SLE. The possibility of a risk haplotype with multiple functional variants cannot be formally rejected.

The mechanisms by which reduced levels of CD226 contribute to the breakdown of immune tolerance are unknown. However, our results suggest that the effect of the autoimmune risk haplotype is specific to T and NKT cells. CD226 is a pleiotropic signalling molecule with a role in the development, co-stimulation, and apoptosis of T and NKT cells (173, 175,
An association between NKT cell deficiency and ADs has been identified and the deficiency of CD226 may be contributing to it (183). Interestingly, the CD226 locus displays signals of positive selection (49, 184). The role of CD226 remains to be elucidated but its association already contributes to the understanding of ADs by revealing NKT cells as important actors in their pathogenesis.
Paper IV: Common and Rare Variants in BLK Lead to Reduced Half-Life of the Protein

BLK (B lymphoid tyrosine kinase) is a tyrosine-kinase of the SRC family of proto-oncogenes involved in B-lymphocyte development, differentiation and signalling (185). BLK is mainly expressed in B cells but it is also present in pancreatic β cells, thymus, hair follicles and salivary ducts (186, 187). In B cells, BLK is activated upon BCR stimulation and, together with LYN and FYN, mediates the transmission of signals critical for the development of pre-B cells, pre-BCR-mediated NFκB activation, as well as activation, growth arrest and apoptosis downstream of the BCR (188). BLK phosphorylates CD79A (Igα) and CD79B (Igβ), which form a complex with the BCR, and the immunoglobulin Fcγ receptors FcγRIIA, FcγRIIB and FcγRIIC (189).

The BLK gene was associated with SLE in a GWAS (54). The SLE-risk allele rs13277113-A is located in the gene promoter and correlated with decreased expression of BLK mRNA in B lymphoblastoid cell lines (54). However, the mechanism by which this SNP, or another SNP in LD, affects gene expression had not been characterized yet. Therefore, in Paper III (178), we performed a fine mapping of the BLK gene, aided by imputation and analysis of epigenetic marks associated with the risk variants.

Analysis of the gene revealed long-range LD across BLK with multiple correlated SNPs associated along the whole gene. Thus, by making use of the sliding window strategy, we delimited six blocks and denominated them B1 to B6. The six haplotype blocks represented the “hills and valleys” of the SNP sliding window analysis, depending on whether they displayed association after correction for multiple testing or not, respectively (Figure 10). The overall strongest single marker signal (rs9986833-A, $P_{CMH-corrected} = 1.05 \times 10^{-4}$, OR 1.38 (1.21 to 1.58)) was located within the first block (B1), which covered BLK’s gene promoter and represented the same associated haplotype tagged by rs13277113-A (54) (Paper IV: online table S1). The multi-marker window analysis revealed a much stronger association to a 1.2 kb window delimited by SNPs rs2248932-rs9329246, located at the end of BLK’s first intron ($P = 1.58 \times 10^{-10}$, omnibus haplotype test) (Figure 10). This window was embedded within B3, a block that displayed moderate LD with B1 (Paper IV, Figure 2). Using data from the ENCODE ChIP-Seq experiments in lymphoblastoid cell lines, we demonstrated that both B1 and B3 were enriched for histone marks of regulatory regions as well as binding sites for the transcription factors NFκBp65 and IRF4 (Paper IV, Figure S2 and Table S3).
Figure 10. SLE-associated SNPs and haplotype windows and ENCODE genomic features of the BLK locus. From top to bottom: $-\log_{10}$ of the p value corrected by false discovery rate for 158 SNPs. Below, the association for haplotype windows of two, three and four SNPs is displayed. The horizontal black line marks the threshold for Bonferroni correction ($p=0.05/323$ windows tested). Only SNP windows contained within block B1 (5’ upstream, exon 1 – UTR – and beginning of intron 1), B3 (end of intron 1) and B5 (exons 4, 5 and 6) were associated after Bonferroni correction.

Association analysis conditioning on rs998683 did not provide evidence in favour of an independent association of SNPs within the window rs2248932-rs9329246, probably due to LD. However, we decided to study this region further as it provided the best multi-marker model of association in B1 and B3, it overlapped with several regulatory histone marks, and had a strong indication for a NFκB binding site, which is known to control the expression of BLK (190). By using ChIP-qPCR in Ig-stimulated human Daudi cells, we confirmed that the p50 and p65 subunits of NFκB bind to this gene region, which is enhanced over time and correlated with the downregulation of BLK mRNA expression (Paper IV, Figure 3A-C).

The conditional analysis did provide evidence in favour of an independent associated haplotype block located within B5 ($P=5.16 \times 10^{-7}$, OR=2.75 (1.80 to 4.19)). The minor alleles of six low-frequency variants (MAF=1.46%) occurred in this haplotype: rs55758736, a non-synonymous SNP in exon 4 resulting in an alanine to threonine substitution at amino acid position 71 (Ala71Thr; GCT>aCT), rs76578315 (chr8:11443817) in intron 4, rs76713287 (chr8:11444706), rs80327577 (chr8:11446229), and
rs116667277 (chr8:11446791) in intron 5, and rs2255227 in intron 6 (Paper IV, Figure 2 and Table S2). The SNP rs numbers have been updated in this text according to dbSNP137.

The SLE-associated allele, 71Thr (rs55758736-A), was present in a haplotype ('F8') causing the autosomal dominant maturity-onset diabetes of the young (MODY11; OMIM 613375) (187). In this study, Borowiec et al determined that the Ala71Thr mutation reduces BLK promoter activity, and in pancreatic β-cell lines it attenuates the enhancing effect of BLK on insulin secretion in response to glucose (187). In addition, the Ala71Thr change was predicted to be damaging by the algorithms implemented in (score=0.99) and MutationTaster (score=0.63) (77, 182). Thus, both in vivo and in silico evidence supported Ala71Thr as a candidate functional variant with a deleterious effect. After treating HEK293T cells transfected with allelic-specific plasmids with the protein synthesis inhibitor cycloheximide, we demonstrated increased protein degradation rates in the presence of the 71Thr mutation.

In Paper IV (178), we demonstrated the association of at least two haplotypes within BLK with SLE, and we provided a possible functional mechanism for the low mRNA and protein levels found in carriers of the risk alleles. The first haplotype contains multiple correlated variants that together non-randomly overlap with TFBSs for NFκB and IRF4. The expression of BLK is known to be under the control of NFκB and the B-cell specific transcription factors PAX5 and EBF1 (190); however, no enrichment for PAX5 or EBF1 binding sites was detected. We hypothesized that multiple sites may explain a coordinated regulatory effect on the expression of BLK. Our study also revealed the independent association of a low-frequency (1.5%) haplotype carrying a deleterious variant (Ala71Thr) that correlated with reduced protein half-life. The addition of a threonine in position 71 creates a putative phosphorylation site that may enhance not only BLK activation but also ubiquitination (191, 192), which may be responsible for the accelerated protein degradation observed in our study.

To date, polymorphisms in the BLK promoter region have been associated with SLE (54, 57), RA (193, 194), PAPS (160), SSc (195), and pSS (196). Our study contributes to the understanding of the mechanisms by which the SLE risk alleles lead to a reduced level of BLK, both at the transcript and at the protein level. However, the mechanism by which reduced levels of BLK leads to increased susceptibility to ADs is not yet understood. Functional studies in native B and T cell types determined that the allele rs922483-A, which is in LD with rs998683-A (r²=0.782 in the CEU population, HapMap 3 release 2), is prominently associated with the reduction of BLK transcript and protein levels in primary human transitional and naïve B cells isolated from umbilical cord blood, but no effect was detected in B cells isolated
from neither adult peripheral blood or tonsils (197). As the former subsets represent more primitive and immature populations of B cells, the authors suggested that low BLK expression predisposes to autoimmunity by a mechanism likely mediated early in B cell development (197). In mice, the triple deficiency of Blk, Lyn and Fyn negatively affects NFκB induction and B-cell development. However, not in the absence of Blk alone, reflecting the redundancy of the Src kinases (198). However, the single deficiency of murine Blk causes reduction of B220+ cells (199), impaired development and hyper responsiveness of MZ B cells after BCR stimulation, and correlates with increased titres of anti-dsDNA antibodies (198). In chronic myeloid leukaemia (CML), BLK acts as a tumour suppressor gene by inhibiting the proliferation of CML stem cells but not normal stem cells (199). In a murine model of CML, the down regulation of Blk is mediated by Pax5, and Blk deficiency (-/-) results in accelerated disease phenotype as well as higher percentage and number of leukaemia cells in peripheral blood and spleen (199). Treatment with IFNa and infection with EBV also down regulate the expression of BLK (200). Hence, the mechanisms responsible for the down regulation of BLK, as well as its outcome, may depend on the cell type and the cell developmental stage. It is also important to note that the activation of B cells is a dynamic process and, as demonstrated in Paper IV, the levels of BLK transcript and protein decrease over time after BCR activation in cells carrying the risk allele.

In summary, BLK appears to be a negative regulator of B-cell reactivity to BCR stimulation and its deficiency increases the susceptibility to pathogenic autoimmunity. Lower levels of BLK, determined by an inherited risk haplotype, exposure to high amounts of interferon, and infection with EBV altogether might therefore exert a cumulative effect on risk through the negative regulation of BLK in B cells (70).
Paper V: Statistical and Functional Epistasis of the B-cell Genes BANK1 and BLK

In Paper V (44), we demonstrated the statistical and functional epistasis between BANK1 and BLK, two established susceptibility genes for SLE (44, 75, 136, 160, 201). The effect of genetic factors that function primarily through a complex mechanism that involves multiple other genes, like BANK1 (B-cell scaffold protein with ankyrin repeats 1), might be missed or underestimated if the gene is examined in isolation. Therefore, epistatic association studies are justified in order to finely dissect the complex genetic architecture of human traits. Epistasis may not only explain larger risk effects than genetic effects alone, but it also serves as a method for revealing unknown and/or non-obvious biological and biochemical pathways that underlie disease pathogenesis.

BANK1 is a scaffold protein that interacts with Src kinases during BCR signalling. Upon BCR stimulation, it physically associates with LYN and to the inositol triphosphate receptors type 1 (IP3R-1) and IP3R-2 promoting their LYN-mediated phosphorylation, which in turn induces Ca\(^{2+}\) mobilization from endoplasmic reticulum stores (202). BANK1 is also phosphorylated by Src kinases, predominantly by SYK (202). Interaction with BLK was not known, thus we hypothesized that BANK1 and BLK interact on the susceptibility to develop SLE and tested whether they also interact at the molecular level in B cells.

Three disease-associated SNPs have been described in BANK1 (4q24): rs10516487 (R61H) located in exon 2; rs17266594, which is a proxy of rs10516487 (\(r^2 = 0.90\)) and is located in a branch point site in intron 2; and an independent signal in exon 7 coding for an ankyrin domain (rs3733197, A383T) (44). The risk allele of rs17266594 alters the splicing efficiency of exon 2 leading to higher levels of the full-length BANK1 isoform and lower levels of an alternative isoform lacking exon 2 (Δ2 isoform) (44). Exon 2 encodes a putative binding domain for IP3R-2, which is known to associate with BANK1 upon B-cell activation.

We extracted genotype data from a 100,000 SNPs-GWAS conducted in 279 cases with SLE and 515 controls from northern Europe (44). Nine tag SNPs in BANK1 (rs7675129, rs10516487, rs10516483, rs2850390, rs1872701, rs10516490, rs1395306, rs871153 and rs238486) were tested for interaction against seven tag SNPs in BLK (rs1478895, rs1478890, rs2252534, rs1382566, rs9329246, rs7014565 and rs2061830) using the algorithm implemented in the GPAT16 method by Wirapati et al (126)(Figure 5). Two SNPs in BANK1 (rs10516483 and rs10516487, \(D'=0.86, r^2=0.36\)) and two in
BLK (rs1478895 and rs2736340, $D'=0.93$, $r^2=0.06$) were involved in significant epistatic interactions and were thus chosen for replication in two independent sets of European-derived populations referred to as “USA” and “Europe” in Paper V (Paper V, Table 1). The strongest association was displayed by the combination of a recessive model for BANK1 rs10516487 (GG) and a dominant model for of BLK rs2736340 (TT+TC) ($P_{\text{meta-analysis}}=1.75 \times 10^{-15}$, OR 1.52 (1.31-1.76)). Regarding the epistatic interactions, the GPAT16 method does not allow yet to introduce covariates or to adjust by different strata. Thus, in order to avoid bias introduced by the different origin of the cohorts, we analysed them separately even though this implied a sacrifice in statistical power. The epistatic interaction of BLK rs2736340 (TT+TC) with BANK1 rs10516487 (GG) ($P_e=0.0013$) and BANK1 rs10516483 (CC) ($P_e=0.0024$) were replicated in the “Europe” data set (Paper V, Table 1).

At the molecular level, we demonstrated the co-immunoprecipitation of BANK1 and BLK proteins in HEK293T cells co-transfected with plasmids expressing BANK1 and BLK, between the native proteins expressed in a B-cell Daudi cell line, and between the native proteins expressed in primary naïve B-cells (Paper V, Figure 1). The binding was enhanced after BCR-stimulation with anti-IgM suggesting that the interaction BANK1-BLK is triggered by B-cell activation. By confocal microscopy and immunostaining, we could also demonstrate that BANK1 alters the subcellular localization of BLK retaining it in the cytoplasm (Paper V, Figure 2).

The detection and replication of interactions is still a challenge and a subject of debate (87, 203). Several methods are available and multiple definitions of epistasis are used (87). We provided in Paper V a functional validation of the GPAT16 method developed by Wirapati et al, which is based on a similar principle to the multifactor dimensionality reduction (MDR) and slightly more powerful than standard algorithms (126). Moreover, using the 100,000 SNPs-GWAS data and the same method, we tested whether genetic interactions involving BANK1 at the genome-wide level could reveal biologically relevant interactions (unpublished results). BANK1 statistically interacted with 27 genes, including the described epistasis with BLK. Among the other genes was ITPR2, which encodes for IP3R-2 and is known to associate with BANK1 upon B cell activation (202). One-fifth of the SLE patients (21%) were homozygous for the risk genotypes of the three genes BANK1, BLK and ITPR2 as compared to 8% of controls (OR=3.20 (2.04-5.01) $P_e<0.0002$).

Our approach provides a system in which the detection of epistatic interactions combined with functional validation can identify pathways of relevance in SLE. We confirm that information about all the molecular
interactions within a cell may be used to reduce the biological hypothesis space for testing for epistasis, as suggested by Sun et al (88).

The pathway involving \textit{BANK1}, \textit{BLK} and possibly \textit{ITPR2} may have important consequences on B cell responses to self-antigens in human autoimmunity (\textit{Figure 11}). Unfortunately, the translation of statistical epistasis to functional epistasis is not straightforward. Our data suggest that \textit{BANK1} may function to sequester \textit{BLK} (and possibly other Src kinases) from the BCR to restrict a sustained BCR signalling. To understand how the particular variants in each gene interact to increase the susceptibility to SLE is even a greater challenge. The risk haplotypes of \textit{BLK} reduce the levels of the transcript and the protein (Paper IV)\textsuperscript{(204)} whereas the risk haplotype of \textit{BANK1} is correlated with higher levels of the full-length isoform \textsuperscript{(44)}. The BCR-mediated signalling may be impaired due to alteration of the balance between \textit{BANK1} and \textit{BLK} concentrations. In scaffolded protein kinase cascades, the deficiency as well as the excess of the scaffold relative to the kinases that it binds results in a decrease of the output of the cascade, which is better known as combinatorial inhibition \textsuperscript{(205)}. This may explain the impairment of the BCR signalling mediated by the deficiency of \textit{BLK} and the excess of \textit{BANK1}. 
Figure 11. Functional model of gene expressed in B-cells and associated with autoimmune diseases in human. Molecular trafficking and sequential protein interactions are major mechanisms to regulate the immune response. Extracellular cues and immune complexes are taken up through the BCR and FcR immune receptors. Nucleic acids and immune complexes are internalized and delivered to intracellular compartments. BANK1 and other SLE-associated genes act in the integration of these signals that leads to production of proinflammatory cytokines, immunoglobulins or type I interferon. The regulation of this pathway is essential for maintain B cell tolerance. BANK1 also interacts with kinases associated with the BCR and the molecular switch PLCγ2. This pathway connects the antigen receptor with extracellular grown factors receptors i.e. the BAFF-receptor that ultimately will dictate cellular decisions such differentiation, growth, and survival. Susceptibility genes are circled in red. FCγR, Low affinity immunoglobulin gamma Fc region receptor; PTPN22, protein tyrosine phosphatase, non-receptor type 22; BAFF, B-cell activating factor; RASGRP3, RAS guanyl releasing protein 3 (calcium and DAG-regulated); LYN, v-yes-1 Yamaguchi sarcoma viral related oncogene homolog; BLK, B lymphoid tyrosine kinase; A20 or TNFAIP3, tumor necrosis factor, alpha-induced protein 3; TRAF6, TNF receptor-associated factor 6; UBE2L3, ubiquitin-conjugating enzyme E2L 3; IRAK1, interleukin-1 receptor-associated kinase 1; TNIP1, TNFAIP3 interacting protein 1; IRF5, interferon regulatory factor 5. Courtesy of Casimiro Castillejo-López.
Conclusions

This thesis includes case-control association studies where the role of the genes TNFSF4 (Paper I), STAT4 (Paper II), CD226 (Paper III), and BLK (Papers IV and V) in the susceptibility of developing SLE was investigated. The primary focus was on the identification of the functional variants underlying the association.

In Paper I, the association of TNFSF4 with SLE was validated and attributed to a risk haplotype tagged by SNPs rs1234317-T and rs12039904-T. Paper II provides evidence supporting the presence of at least two independent genetic effects within the STAT4 gene represented by rs3821236-A and rs7574865-A, which correlated with increased levels of gene expression. These variants act additively with IRF5 on the risk to develop SLE. In Paper III, a functional allele in CD226 (rs727088-C) was identified, which was responsible for the decreased levels in both mRNA and protein expression. In Paper IV, two independent genetic effects in the BLK gene were demonstrated. The first one comprised multiple regulatory variants in high LD that were enriched for NFκB and IRF4 binding sites and correlated with low BLK mRNA levels. The second was a low-frequency missense substitution (Ala71Thr) that decreased the half-life of the BLK protein. In Paper V, a genetic epistatic interaction between BANK1 rs10516487 (GG) and BLK rs2736340 (TT+TC) was demonstrated. Additional molecular analyses established that these molecules interact physically.

These studies have contributed to the dissection of the genetic architecture of SLE. They highlight the allelic heterogeneity of the disease, which may involve both common and rare deleterious variants within the same locus. This allelic heterogeneity may reflect differences in genetic ancestry and clinical presentation. Furthermore, we provide evidence supporting the contribution of gene-gene epistatic interactions on the disease susceptibility. These studies also reveal the functional role of several SLE-associated variants, which has significantly aided in the understanding of SLE disease pathogenesis.
Future Perspectives

Since this doctoral thesis started in 2007, the study design strategies, the methods used, and the whole way of thinking have inevitably been influenced by the increasing knowledge about the human genome, the explosion of methods to characterize the genetic diversity among individuals and populations, and the publicly available resources. The papers included in this thesis and related publications are part of an era of studies that have significantly accelerated the understanding of the genetic architecture of SLE and ADs. The ability to collect genetic information rapidly and inexpensively, including sequence and genotypic variation, has not been accompanied by an equal development in the methods used to study gene and variant function in vitro and in vivo. Efforts like the ENCODE Project have generated invaluable resources that now allow the functional study of the human genome in a comprehensive way, for particular traits or cells of interest, in the same way as GWAS provided a systemic method to map genes associated with diseases and quantitative traits.

The identification of nearly 50 SLE susceptibility loci has elucidated important pathogenic pathways underlying disease development. Without functional links to the associated variants, it is not possible to translate genetic discoveries into the improvement of patient care. The dissection of functional variants remains to be a bottleneck when translating genetic findings to the clinic and, therefore, deserve more attention. A deep understanding of the disease mechanisms, of the pathways involved and the major hubs controlling them, of the specific effect of the susceptibility variants and, importantly, of the way in that they interact with each other, are all indispensable steps to be able to develop novel therapies and replace the often harmful treatments currently available. If we can fully characterize a pathogenic mechanism and reproduce it in cells or animal models, maybe we will be able to modulate it in patients as well. Immunologic and genetic discoveries are already generating potential therapeutic strategies (19). Belimumab, a human monoclonal antibody against the cytokine BAFF (also known as BLyS) is the first biologic drug approved by the FDA for the treatment of SLE. However, the initial optimism generated by its primary clinical efficacy has been shadowed by the failure to sustain it after 76 weeks of treatment. The mechanisms affecting the limited efficacy are unknown (19).
The differential association of some genes with the presence of particular clinical manifestations and antibodies opens the question of whether we have been dealing with different diseases, all under the “SLE” definition. Different SLE autoantibodies are also associated with specific although non-exclusive clinical phenotypes. It may be that, for example, patients with an early disease onset, glomerulonephritis, and in general a more severe disease are enriched for risk variants. Hence, they could constitute a distinct and/or more “genetic” disease. Alternatively, some traits such as glomerulonephritis may be more “heritable” than others. Given the severity and lethality of lupus nephritis, a set of genetic biomarkers that could predict which patients are more likely to develop this complication would dramatically save a lot of pain and deaths. Genetic biomarkers may also predict which patients would be benefit of certain therapies.

A more difficult goal is to use genetic markers to predict which individuals will develop SLE. Despite the undeniable success of GWAS in SLE, the susceptibility genes explain only a small fraction of the disease heritability, which further raises more questions. For example, are there many yet undiscovered common variants? Are SNPs and the common variation they represent responsible for only a fraction of the heritability and then other types of human polymorphisms, for instance CNVs, explaining the rest? Do we have to increase the sample size to the order of hundreds of thousands of individuals to be able to detect the rest of the variants that account for the “missing” heritability? To be able to predict the disease in a sensitive and specific way, do all the variants behind SLE susceptibility need to be known and tested? This does not seem to be an attainable goal for now.

To define the role of rare and highly penetrant variants in SLE is a challenge. A more pertinent question at present is how genetic variants (common, low-frequency or rare) interact with each other, and how environmental factors act on genetically susceptible individuals. Maybe we have to start by giving diseases the right name to be able to dissect the right genes and the right mechanisms underpinning self-tolerance, and eventually, generate well-targeted treatment. SLE seems to be a collection of different diseases; the ACR criteria are in principle classification and not diagnostic criteria that aimed to define more homogeneous and comparable disease subsets. In addition, symptoms of and other ADs often overlap in the same individual (i.e. SLE and pSS) and hence co-morbidities need to be considered for the classification of patients or, at least, accounted for in any genetic analysis.

We have many pieces of the puzzle in our hands, and also a constantly improving scaffold: the map of the human genome. We still need to learn how to arrange them correctly before we can generate an accurate picture of the genetic basis of SLE.
Resumen en Español

El Lupus Eritematoso Sistémico (LES) es una enfermedad autoinmune sistémica, crónica y a menudo severa, que afecta principalmente a mujeres en edad reproductiva. El LES se caracteriza por la producción de autoanticuerpos contra ácidos nucleicos y sus proteínas asociadas. Estos anticuerpos se acumulan en múltiples órganos y tejidos causando inflamación y daño irreversible. Los ácidos nucleicos están presentes en todas las células del organismo; por esto, cualquier órgano es blanco potencial de la respuesta autoinmune haciendo que las manifestaciones del LES sean altamente variables de un paciente a otro, e incluso en el mismo paciente durante el curso de la enfermedad. Estas manifestaciones abarcan desde el típico “eritema en mariposa” en la piel hasta nefritis severa e insuficiencia renal, siendo esta última una de las principales causas de mortalidad en pacientes jóvenes.

Aún se desconocen los mecanismos patológicos que llevan a la pérdida de la tolerancia inmunológica pero se sabe que dependen tanto de factores genéticos como no-genéticos (i.e. ambientales y hormonales). Desde el punto de vista genético, el LES es una enfermedad compleja, lo que quiere decir que es poligénica. La primera evidencia de un componente genético viene de estudios familiares que han demostrado un incremento en la prevalencia de la enfermedad en hermanos de pacientes con LES ($\lambda_s=8–29$), fenómeno conocido como agregación familiar. Además, la tasa de concordancia de la enfermedad para gemelos monocigotos (20-40%), los cuales comparten el 100% de su material genético, es 10 veces más alta que para gemelos dicigotos y otros hermanos (2-5%), los cuales comparten solo la mitad. Estos estudios han estimado que la heredabilidad de la enfermedad es del 66%.

Diferentes estrategias se han usado para revelar la localización de los genes de susceptibilidad para LES en el genoma humano. Mediante la tipificación de millones de polimorfismos de un solo nucleótido (SNPs) distribuidos a lo largo del genoma, hoy es posible realizar mapeos de asociación a nivel genómico o GWAS, por su sigla en inglés. Los estudios de asociación genética investigan si existe correlación estadística entre una o más variantes genéticas y un rasgo dado, a partir de como difieran las frecuencias de estas variantes entre un grupo de pacientes (casos) y otro de
individuos sanos que provienen de la misma población y que tienen una distribución similar de edad y sexo (controles).

Múltiples genes se han asociado con LES (resumidos en la Tabla 2). Sin embargo, para la mayoría de estos genes, aún se desconoce cuáles son las variantes genéticas directamente asociadas con la enfermedad, así como sus consecuencias en la función de estos genes o en las proteínas que estos codifican. Los genes de susceptibilidad son de alguna manera las piezas de un rompecabezas, su identificación es el primer paso pero no podemos armar el rompecabezas si no entendemos en qué orden están arregladas las piezas, qué pasa si una falta y cómo interactúan entre ellas. Una vez podamos entender los eventos moleculares que contribuyen al LES, tal vez seamos capaces de intervenirlos y traducir los resultados de este tipo de investigaciones en el desarrollo de nuevos tratamientos.

Esta tesis doctoral incluye cinco estudios de asociación que investigan el papel de los genes de susceptibilidad *TNFSF4* (Paper I), *STAT4* (Paper II), *CD226* (Paper III), y *BLK* (Papers IV & V) en la susceptibilidad a desarrollar LES. El objetivo principal de estos estudios fue la identificación de variantes funcionales. El primer estudio (Paper I) valida la asociación del gen *TNFSF4* y además demuestra que, en Europeos, el haplotipo de riesgo está definido por la presencia de los alelos en dos SNPs: rs1234317-T y rs12039904-T. En el segundo estudio (Paper II) se presenta evidencia a favor de dos efectos genéticos independientes en el gen *STAT4*, representados por los SNPs rs3821236-A y rs7574865-A. Además, los resultados demuestran que la expresión génica de *STAT4* está incrementada en individuos portadores de estos alelos de riesgo. El tercer estudio (Paper III) revela una variante funcional en el gen *CD226* (rs727088-C), la cual es responsable de una reducción en la expresión del gen y en los niveles de la proteína, específicamente en linfocitos T y células natural killer (NK) T. En el cuarto estudio (Paper IV) se demuestra la presencia de al menos dos efectos genéticos independientes en el gen *BLK*. El primer efecto genético involucra múltiples alelos de riesgo presentes en el mismo haplotipo, los cuales están localizados en los sitios de unión al ADN de los factores de transcripción NFκB e IRF4 y se asocian a bajos niveles de expresión del gen. El segundo efecto genético depende de una variable no sinónima que cambia una alanina por una treonina en la posición 71 de la proteína (A71T), reduciendo su vida media. El último estudio (Paper V) revela que las variantes de susceptibilidad en los genes *BLK* y *BANK1* interactúan de forma no aditiva en la susceptibilidad a desarrollar LES. Es decir, el efecto conjunto de los dos genes es mucho mayor que el esperado por la suma de sus efectos individuales. Además se demuestra en este estudio, por primera vez, que las proteínas codificadas por estos genes interactúan físicamente en el citoplasma de linfocitos B.
Estos estudios contribuyen al entendimiento de la genética del LES. Por un lado, resaltan que pueden existir múltiples variantes genéticas dentro del mismo gen asociadas con la enfermedad, fenómeno conocido como heterogeneidad alélica. Esta heterogeneidad alélica puede involucrar alelos comunes que tienen un efecto leve en la función del gen y la proteína, así como variables más raras con un efecto deletéreo. Diferentes alelos en el mismo gen pueden estar presentes en individuos con diferente ancestría genética o con diferentes manifestaciones de la enfermedad. Adicionalmente, en esta tesis se demuestra que existe interacción genética o epistasis entre algunos genes de susceptibilidad al LES y la contribución de estas interacciones a nivel genómico merece ser estudiada en detalle. En general, estos estudios revelan variantes funcionales que afectan la regulación de la expresión de los genes \textit{STAT4}, \textit{CD226} y \textit{BLK} así como los niveles de las proteínas que estos codifican. Los resultados de esta investigación contribuyen significativamente al esclarecimiento de los mecanismos genéticos y patogénicos que conllevan a la pérdida de la tolerancia inmunológica y al desarrollo de LES.
I am and will always be immensely grateful to my supervisor, Professor Marta E. Alarcón Riquelme, for trusting me from the moment we met at a workshop in Cuba when I was still a Master student in Colombia. Without knowing me, you believed in me and brought me to Sweden. I hope I did not disappoint you (I never made it to Nature Genetics, but how we tried!). I appreciate all you have taught me, the place you made for me in your research, and how you encouraged me to learn and make the best out of this PhD. Thank you, Marta, for being my mentor. I wish you the best in all the new adventures and I am sure you will continue succeeding because you are intelligent, creative, have true passion for what you do, and fight for it like an authentic Mexican warrior. You will always have my admiration and friendship.

I am also very grateful to Professor Ulf Gyllensten and his group, for “adopting” me when Marta had to move to Spain. I truly enjoyed all the academic and non-academic activities. Ulf, you came in the right moment to give me the structure and the clarity to sort out the priorities during the final phase of my PhD. It has never been so easy for me to finish something, thanks to you everything was smooth and there was no (much) stress. Inger Jonasson, thanks for that hug! You don’t have idea how much I needed it that day.

This thesis was possible thanks to a great team job. I am indebted to my colleagues and friends from the disappeared Genetics of Autoimmune Diseases group: Sergey Kozyrev, Casimiro Castillejo, Hong Yin, Anna Karin Abelson, Serena Fineschi, Juan Ramón López, Ammar Zaghool, Sara Löfgren and Caroline Gallant. Thanks for all the scientific discussions and the moments shared outside the lab. Ammar, I will never forget our conference in Hawaii and all the privileges we enjoyed when people thought we were honeymooners 😊 To Sara and Caroline, we were united through motherhood and you have a very special place in my heart.

Thank you, to all students and guest researchers that came to the lab and contributed to this work: Sandra Sveder, Sahar Ardabili, Alexandra Bozikova, Michaela Fredi, Lina Marcela Diaz, Liliya Sungatullina, and Elnara Biktagirova. To all that passed by the SLE group, Hanna, Juliana, David P, Sara C and many more that I am surely missing, big Thank You!
I would like to give special thanks to our collaborators. To all the clinical researchers from the BIOLUPUS group, who collected the samples and clinical data, you play a key role in this type of research. I would like to give special thanks to Jérôme Wojcik for inviting me to Merck Serono in Geneva and teaching me the epistatic interaction method. Thank you also for being so helpful during all the submission and re-re-submission process! I am also especially thankful to Professor David B. Goldstein for hosting me in his lab at Duke University and for the fruitful discussions that shaped my view of the genetic architecture of SLE.

I thank everybody at the Department of Genetics and Pathology at Uppsala University for providing such a nice work environment. To the administration personal that kindly helped me with my travel grants, visa applications, and more, always smiling. To all the PhD students that became my hearty friends, party mates, and/or PhD-panic fellows! Emma, Sara, Mi, Jelena, Nikki, Demet, Paul, Fiona, Lesley. To my office-mates, Maria, Viktoria, Joaquim, & Stephanie, who made of my last days at IGP a very enjoyable time (and also made me eat a lot of sushi). I might have lost my SLE group, but I was never alone!

It is impossible to unlink my PhD to my life in Sweden because they started at the same time. I came almost six years ago, full of dreams and motivation to work but also overwhelmed by the sadness of living my family and my friends in Colombia. I feel the luckiest person in the world because in Sweden I found more than a job. Sweden gave me my prince 😊, my princess, a family and wonderful friends.

To the T-shirts club: Tanai, Daniel OD, Caro, Ana, Jessica, Daniel C, Tobi, Anna E, Ola, Claire: It’s been so easy to love you. Thank you for every single moment we have shared together. Thank you for coming to Colombia and making of our wedding an unforgettable memory. I can’t wait to see you all again!

Gracias Leonor, Rolf & familia, y a Luz Stella & Lars, por convertirse en mi familia de corazón en Suecia.

Tack till alla underbara människor i Tobias familj för att ni har gett mig en fin plats i ert liv. Tack framför allt till Ann!

Tobias, there are no words that can express how much I love you and how lucky I feel to have you. I immensely appreciate all the sacrifices that you have made to help me with my PhD, I truly do. Thank you for doing everything at home so that I could focus on writing, for going to Rudbeck to bring me a computer charger in the middle of the freezing night, for showing me that love and equality is more about doing everything for each other. I’d do everything and anything for you and Emilia.
Mi Emi, gracias por llenarme la vida y darme fuerza para trabajar sin descanso. Te amo con todo mi corazón y espero que olvides la madre inatenta que has tenido en los últimos días.

A mi familia, gracias por mantenerte cerca de mí a pesar de la distancia. Cuánto quisiera poder abrazarlos cada vez que los extraño, es decir, todos los días! Hoy termino mi doctorado gracias al apoyo incondicional que siempre he recibido de ustedes. Gracias especialmente a mi mamá, por poner en mis manos el primer documento que leí sobre el Proyecto Genoma Humano cuando apenas tenía 15 años, siempre alimentando mi curiosidad y mis ambiciones. No me alcanza la vida para devolverte el inmenso sacrificio que hiciste para que yo pudiera estudiar medicina y sacar adelante mi maestría. Te debo todo madrecita.

Last but not least, I am indebted to Jess and Caroline, for proofreading this thesis, and to Daniel OD, for drawing the gorgeous cover. The butterfly has become the international symbol of SLE, so I wanted a Colombian endemic butterfly on my cover (*Morpho peleides*). But you gave me the whole country in it, the emeralds, the rivers, and the mountains. I love it!

Gracias a Dios. Gracias a todos.
References


79


A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine.