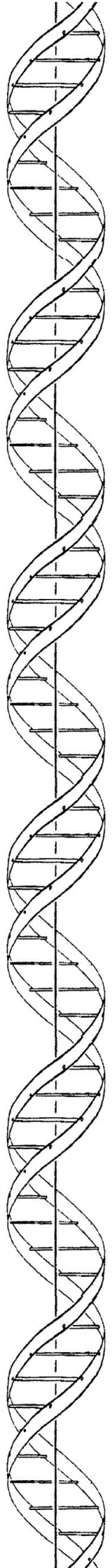


Tesis Doctoral

Systemic Sclerosis and the Genetic Continuum

Memoria presentada por el licenciado en biología
José Ezequiel Martín Rodríguez para optar al grado
de Doctor Internacional por la Universidad de
Granada.

Director: Javier Martín Ibáñez, Profesor de
Investigación del CSIC.



CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

CSIC



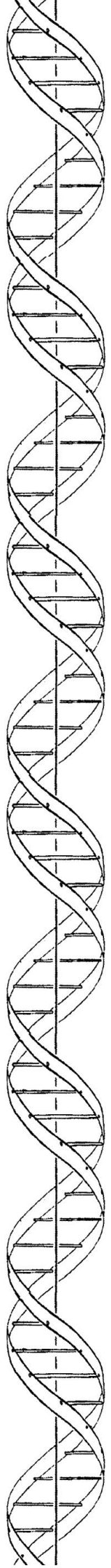
ugr

Universidad
de **Granada**

Instituto de Parasitología y Biomedicina López-Neyra, CSIC.

Granada, Diciembre de 2012.

Editor: Editorial de la Universidad de Granada
Autor: José Ezequiel Martín Rodríguez
D.L.: GR 1724-2013
ISBN: 978-84-9028-562-6



El doctorando José Ezequiel Martín Rodríguez y el director de la tesis Javier Martín Ibáñez garantizamos, al firmar esta tesis doctoral, que el trabajo ha sido realizado por el doctorando bajo la dirección de los directores de la tesis y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

Así mismo, el doctorando José Ezequiel Martín Rodríguez y el director de la tesis Javier Martín Ibáñez garantizamos, al firmar esta tesis doctoral, que las cinco publicaciones presentadas no se han utilizado en la defensa de ninguna otra tesis en España u otro país y que no serán utilizadas con tal propósito.

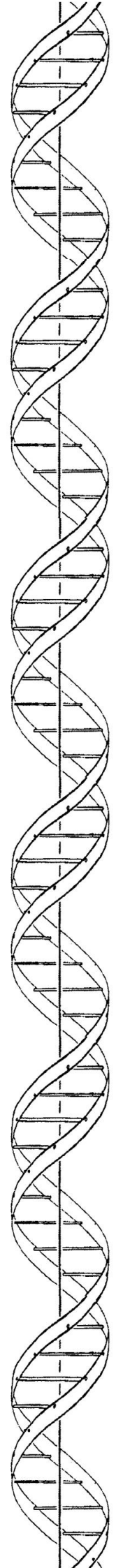
Granada, 18 de Diciembre de 2012

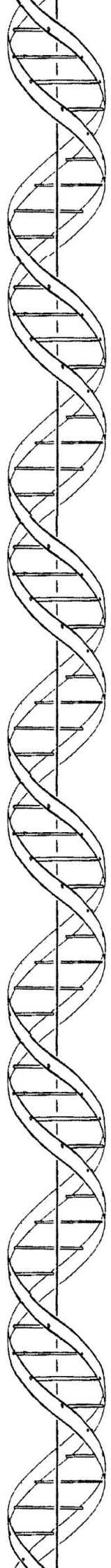
Director de la Tesis

Doctorando

Fdo.: Javier Martín Ibáñez

Fdo.: José Ezequiel Martín Rodríguez





From the Oxford dictionary:

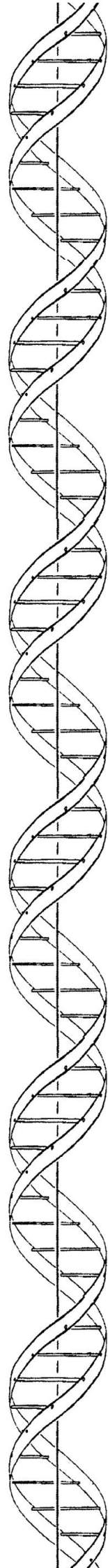
Genetic: /dʒɪ'netɪk/ relating to genes or heredity.

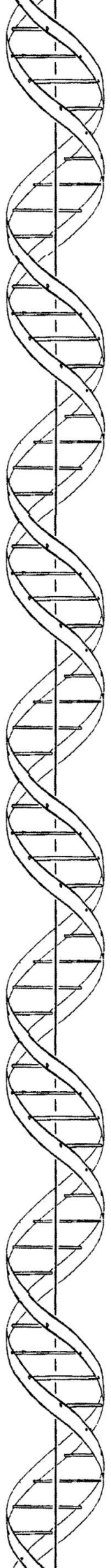
Gene: /dʒi:n/ a unit of heredity which is transferred from a parent to offspring and is held to determine some characteristic of the offspring.

Continuum: /kən'tɪnjʊəm/ a continuous sequence in which adjacent elements are not perceptibly different from each other, but the extremes are quite distinct.

INDEX

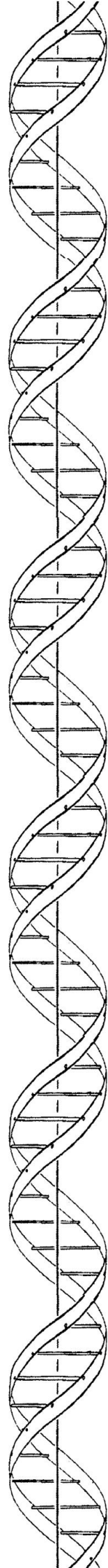
Abbreviations	4
Summary	8
Introduction	14
Systemic Sclerosis	16
<i>Environmental Factors in Systemic Sclerosis</i>	20
<i>Genetic Component of Systemic Sclerosis</i>	22
The Advent of the Genomic Era	26
Objectives	30
Publications	34
Genome-wide association study of systemic sclerosis identifies <i>CD247</i> as a new susceptibility locus. <i>Nature Genetics</i> , 2010.....	36
Identification of Novel Genetic Markers Associated with Clinical Phenotypes of Systemic Sclerosis through a Genome-Wide Association Strategy. <i>PLoS Genetics</i> , 2011.	58
Identification of <i>CSK</i> as a systemic sclerosis genetic risk factor through Genome Wide Association Study follow-up. <i>Human Molecular Genetics</i> , 2012.	86
Seven aminoacids in <i>HLA-DRB1</i> and <i>HLA-DPB1</i> explain the majority of MHC associations with systemic sclerosis. Under Review.	102
Systemic sclerosis and systemic lupus erythematosus pan-meta-GWAS reveals six new shared susceptibility loci. Under Review.	134
Discussion	174
New Genetic Findings in Systemic Sclerosis	176
<i>Biological Relevance</i>	179
<i>Discerning Capability</i>	184
Pan-Autoimmunity.....	189
The Genetic Continuum.....	194
Future Directions	199
Conclusions	202
Bibliography	206





IN LABORING TO BE CONCISE, I BECOME OBSCURE.

-HORACE



ABBREVIATIONS

ACA: anti-centromere auto-antibody.

ACPA: Anti-Citrullinated Protein Auto-antibody.

AID: Autoimmune Disease.

ATA: anti-topoisomerase I auto-antibody.

ATG5: autophagy related 5.

BANK1: B-cell scaffold protein with ankyrin repeats 1.

BLK: B lymphoid tyrosine kinase.

CD247: CD247 molecule.

CR: Cumulative Risk.

CSK: c-src tyrosine kinase.

SSc: Systemic Sclerosis, Scleroderma.

dcSSc: diffuse cutaneous subtype of systemic sclerosis.

DNA: Deoxyribonucleic Acid.

FCGR2A: Fc fragment of IgG, low affinity IIa, receptor (CD32).

GO: Gene Ontology.

GRAIL: Genetic Relationships Across Implicated Loci.

GWAS: Genome-wide association study

HLA: Human Leukocyte Antigen.

HLA-DPB1: major histocompatibility complex, class II, DP beta 1.

HLA-DRB1: major histocompatibility complex, class II, DR beta 1.

ICA1: islet cell autoantigen 1, 69kDa.

IKZF1: IKAROS family zinc finger 1.

IL10: interleukin 10.

IL12RB2: interleukin 12 receptor, beta 2.

IL2RA: interleukin 2 receptor, alpha.

IRAK1: interleukin-1 receptor-associated kinase 1.

IRF5: interferon regulatory factor 5.

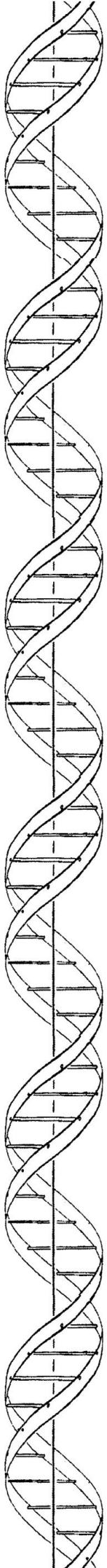
IRF7: interferon regulatory factor 7.

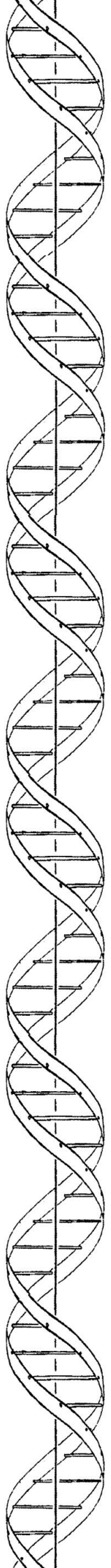
IRF8: interferon regulatory factor 8.

ITGAM: integrin, alpha M (complement component 3 receptor 3 subunit).

ABBREVIATIONS

- JAZF1:** JAZF zinc finger 1.
- KIAA0319L:** KIAA0319-like.
- lcSSc:** limited cutaneous subtype of systemic sclerosis.
- LD:** Linkage Disequilibrium.
- MHC:** Major Histocompatibility Complex.
- MICB:** MHC class I polypeptide-related sequence B.
- NARAC:** North America Rheumatoid Arthritis Consortium.
- NFKB1:** nuclear factor of kappa light polypeptide gene enhancer in B-cells 1.
- NOTCH4:** notch 4.
- OR:** Odds Ratio.
- PSD3:** pleckstrin and Sec7 domain containing 3.
- PTPN22:** protein tyrosine phosphatase, non-receptor type 22 (lymphoid).
- PTTG1:** pituitary tumor-transforming 1.
- PXK:** PX domain containing serine/threonine kinase.
- RA:** Rheumatoid arthritis.
- SAMD9L:** sterile alpha motif domain containing 9-like.
- SLE:** Systemic Lupus Erythematosus.
- SNP:** Single nucleotide polymorphism
- SOX5:** SRY-box containing gene 5.
- STAT4:** signal transducer and activator of transcription 4.
- TNFAIP3:** tumor necrosis factor, alpha-induced protein 3.
- TNFSF4:** tumor necrosis factor (ligand) superfamily, member 4.
- TNIP1:** TNFAIP3 interacting protein 1.
- TYK2:** tyrosine kinase 2.
- UBE2L3:** ubiquitin-conjugating enzyme E2L 3.
- UHRF1BP1:** UHRF1 binding protein 1.
- WTCCC:** Welcome Trust Case Control Consortium.

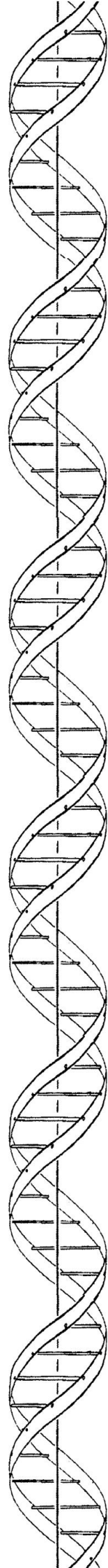




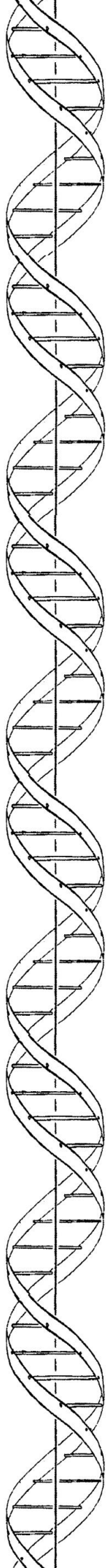
VIGOROUS WRITING IS CONCISE. A SENTENCE SHOULD CONTAIN NO UNNECESSARY WORDS, A PARAGRAPH NO UNNECESSARY SENTENCES, FOR THE SAME REASON THAT A DRAWING SHOULD HAVE NO UNNECESSARY LINES.

-WILLIAM STRUNK

SUMMARY



SUMMARY



Systemic sclerosis (SSc) is a complex heterogeneous disease with a genetic and an environmental component characterized by three main pathological courses: vascular damage, altered immune response and extensive fibrosis of the skin and internal organs which ultimately leads to the death of the patient. SSc presents two major subtypes: the limited cutaneous subtype (lcSSc) and the diffuse cutaneous subtype (dcSSc); and two major auto-antibodies: anti-DNA topoisomerase I (ATA) and anti-centromere auto-antibodies (ACA).

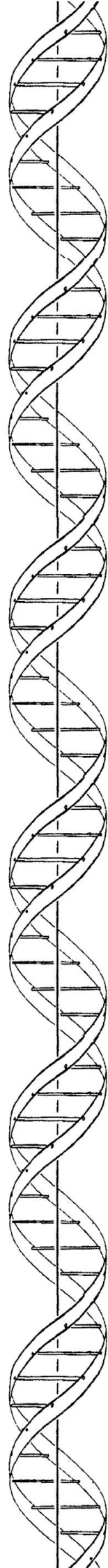
Prior to the beginning of this thesis, few loci involved in SSc were described. In order to extend the knowledge of the genetics of SSc we have performed a genome-wide association study (GWAS) of 2,771 SSc patients and 5,706 controls. The range of approaches used includes the throughout analysis of all the GWAS and suggestive level signals, the analysis of the main subphenotypes of SSc and the pan-meta-analysis of systemic lupus erythematosus and SSc GWAS data, all followed by the replication of findings in follow-up independent cohorts including 3,237 patients and 6,097 controls. Thanks to these, we have been able to identify 13 new SSc susceptibility loci: *ATG5*, *CD247*, *CSK*, *IKZF1*, *IRF8*, *JAZF1*, *KIAA0319L*, *NFKB1*, *NOTCH4*, *PSD3*, *PXK*, *SAMD9L* and *SOX5*.

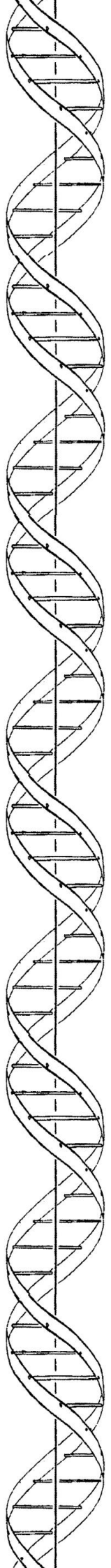
Of these new susceptibility loci, the majority represent functions in different compartments of the immune system like T cell biology (*CD247* and *CSK*), B cell biology (*IKZF1*), autophagy (*ATG5*), inflammation (*PXK* and *NFKB1*) and innate immunity (*IRF8*), but most importantly we have been able to uncover the first two susceptibility loci in which we find genes involved in one of the three major hallmarks of SSc: collagen deposit and fibrosis (*NOTCH4* and *SOX5*).

Additionally, through the imputation of HLA classical alleles and polymorphic aminoacidic positions using GWAS data, we refined the historically well-known peak of association in this region. We have been able to define a set of seven aminoacids in the HLA-DR β 1 and HLA-DP β 1 molecules which explains almost all association observed in the HLA region with SSc and confine it to the ACA and ATA positive subgroups.

Thanks to the throughout analysis of these variants we have gained the knowledge of the compartment in which each of the susceptibility variants belong. Using these data we have learned that the auto-antibody producing subsets of patients (ACA and ATA), are more genetically homogenous entities than the clinically classified groups (SSc, lcSSc and dcSSc).

As a generalization of the discoveries presented in this thesis we propose that **individuals are the combination of many observable phenotypic continuums, which can be subdivided in many other biological continuums, which, in turn, are the product of the interaction between the genetic continuum of the involved loci and the environmental factors.** Thus, individuals grouped under the criteria of biological phenotypes will tend to be more genetically homogeneous than those grouped under clinical classification criteria.





La esclerosis sistémica (SSc) es una enfermedad heterogénea y compleja con componente genético y ambiental que se caracteriza principalmente por tres vías patológicas: daño vascular, respuesta inmune alterada y una extensa fibrosis de la piel y los órganos internos que conduce en última instancia a la muerte del paciente. La SSc presenta dos subtipos principales: el subtipo limitado cutáneo (lcSSc) y el subtipo limitado difuso (dcSSc); y dos auto-anticuerpos principales: auto-anticuerpos anti-DNA topoisomerasa I (ATA) y anti-centrómero (ACA).

Antes de la realización de esta tesis, tan solo unos pocos loci se habían descrito como factores de susceptibilidad para SSc. Para ganar un mayor conocimiento de la genética de la SSc hemos realizado un estudio de asociación del genoma completo (GWAS) en 2,771 pacientes de SSc y 5,706 controles. El abanico de acercamientos utilizados incluyen el meticuloso análisis de todas las señales a nivel de GWAS y a nivel sugestivo, el análisis de los principales subfenotipos de la SSc y el pan-meta-análisis de datos de GWAS de SSc y el lupus eritematoso sistémico, todos ellos seguidos por las correspondientes fases de replicación en cohortes de seguimiento independientes incluyendo otros 3,237 pacientes de SSc y 6,097 controles sanos. Gracias a esto hemos sido capaces de identificar 13 nuevos loci de susceptibilidad a la SSc *ATG5*, *CD247*, *CSK*, *IKZF1*, *IRF8*, *JAZF1*, *KIAA0319L*, *NFKB1*, *NOTCH4*, *PSD3*, *PXK*, *SAMD9L* y *SOX5*.

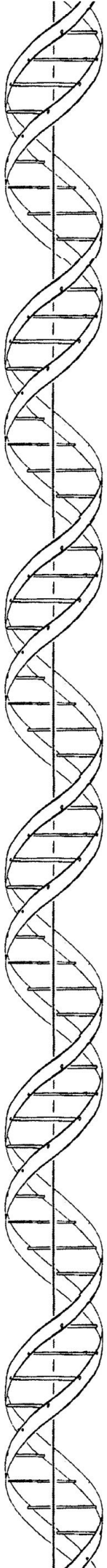
De estos nuevos loci de susceptibilidad, la mayoría representan funciones en diferentes compartimentos del sistema inmune como la biología de las células T (*CD247* y *CSK*), la biología de las células B (*IKZF1*), la autofagia (*ATG5*), la inflamación (*PXK* y *NFKB1*) y la respuesta inmune innata (*IRF8*), pero aún más importante hemos descrito los dos primeros loci de susceptibilidad en los que encontramos genes involucrados en una de las tres principales vías patogénicas de la SSc: el depósito de colágeno y la fibrosis (*NOTCH4* y *SOX5*).

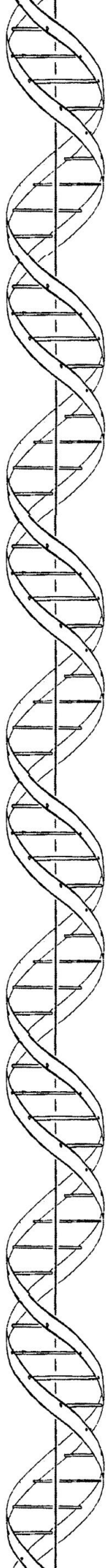
Adicionalmente, a través de la imputación de los alelos clásicos del HLA y sus posiciones aminoácidas polimórficas usando datos de GWAS, hemos refinado la históricamente conocida asociación de esta región. Hemos sido

capaces de definir un conjunto de siete aminoácidos en las moléculas HLA-DR β 1 y HLA-DP β 1 que explican casi toda la asociación observada en la región HLA con la SSc y confinarla a los subgrupos ACA y ATA positivos.

Gracias al meticuloso análisis de estas variantes hemos ganado un mayor conocimiento de a que compartimento pertenece cada asociación de cada uno de los factores genéticos de susceptibilidad. Utilizando estos datos hemos aprendido que los subgrupos positivos para los auto-anticuerpos (ACA y ATA) son entidades más homogéneas genéticamente que los grupos clasificados bajo características clínicas (SSc, lcSSc y dcSSc).

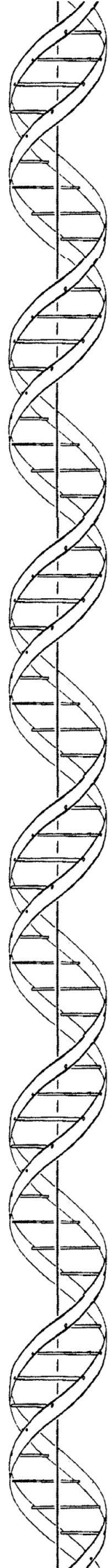
Como una generalización de los descubrimientos presentados en esta tesis proponemos que **los individuos son una combinación de una multitud de continuos fenotípicos observables, los cuales pueden ser subdivididos en muchos otros continuos biológicos, los cuales, a su vez, son el producto de la interacción entre el continuo genético de los loci involucrados y los factores ambientales.** Así, los individuos agrupados bajo criterios de clasificación basados en fenotipos biológicos tenderán a ser más homogéneos genéticamente que aquellos agrupados bajo criterios de clasificación clínica.



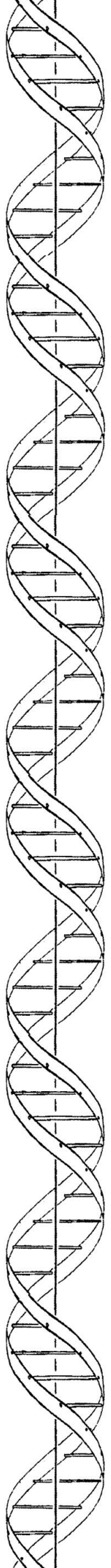


I DON'T WANT TO BELIEVE. I WANT TO KNOW.

-CARL SAGAN



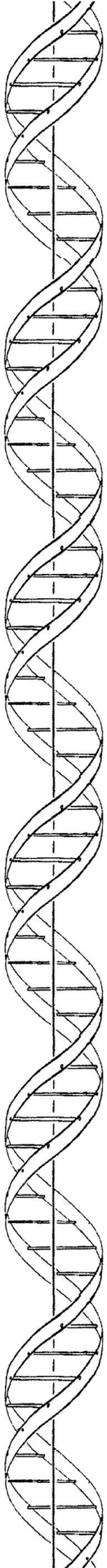
INTRODUCTION



SYSTEMIC SCLEROSIS

Systemic sclerosis or scleroderma (SSc) is a complex heterogeneous disease characterized by three main pathological courses: vascular damage, altered immune response and extensive fibrosis of the skin and internal organs [1]. This disease presents two major clinical subtypes: the limited cutaneous subtype (lcSSc), which is milder and involves usually the fibrosis of skin in the distal parts of the body; and the diffuse cutaneous subtype (dcSSc), which is a more severe form of the disease, progresses much faster and fibrosis affects at least one internal organ in addition to the skin [2]. The vascular damage is the earlier alteration which appears in SSc patients, mainly consisting in the loss of integrity of the endothelial layer and can occur in all organs [3-6]. This vascular damage precedes the fibrosis, as it gradually replaces the vascular inflammatory phase, and ultimately leads to the disruption of the architecture of the affected tissue. This fibrosis is the cause of the main symptoms of the disease, and in the later, more severe stages of the disease is mainly due to the accumulation of type I collagen, especially in the lungs of dcSSc patients [7, 8]. Thus, fibrosis is the ultimate responsible for most complications and death of the patients. The most commonly affected organs by fibrosis in these patients, especially in dcSSc patients, are the lungs [1]. Additionally to the lung fibrosis, and presenting itself as the single major SSc complication which most frequently leads to the death of the lcSSc patients, is the pulmonary arterial hypertension [9]. Pulmonary arterial hypertension develops in up to 30% of patients with SSc, being more frequent in the lcSSc subtype and sometimes overlapping with pulmonary fibrosis [9].

Another of the major features of SSc is the altered immune response, leading to the production of auto-antibodies. Among these auto-antibodies we can find the DNA topoisomerase I (ATA), the anti-centromere auto-antibodies (CENP A and/or B proteins) (ACA), RNA polymerase III (pol-III), U3-RNP (fibrillarin), Th/To, PM/SCL, and U1-anti-ribonucleoprotein (RNP) [1, 10-14]. The production of these auto-antibodies have been described to partially overlap with clinical subtypes and manifestations of the disease, *e.g.* ACA production has been associated with the lcSSc subtype and pulmonary arterial hypertension, while ATA production has been associated with the dcSSc subtype and pulmonary fibrosis [15].



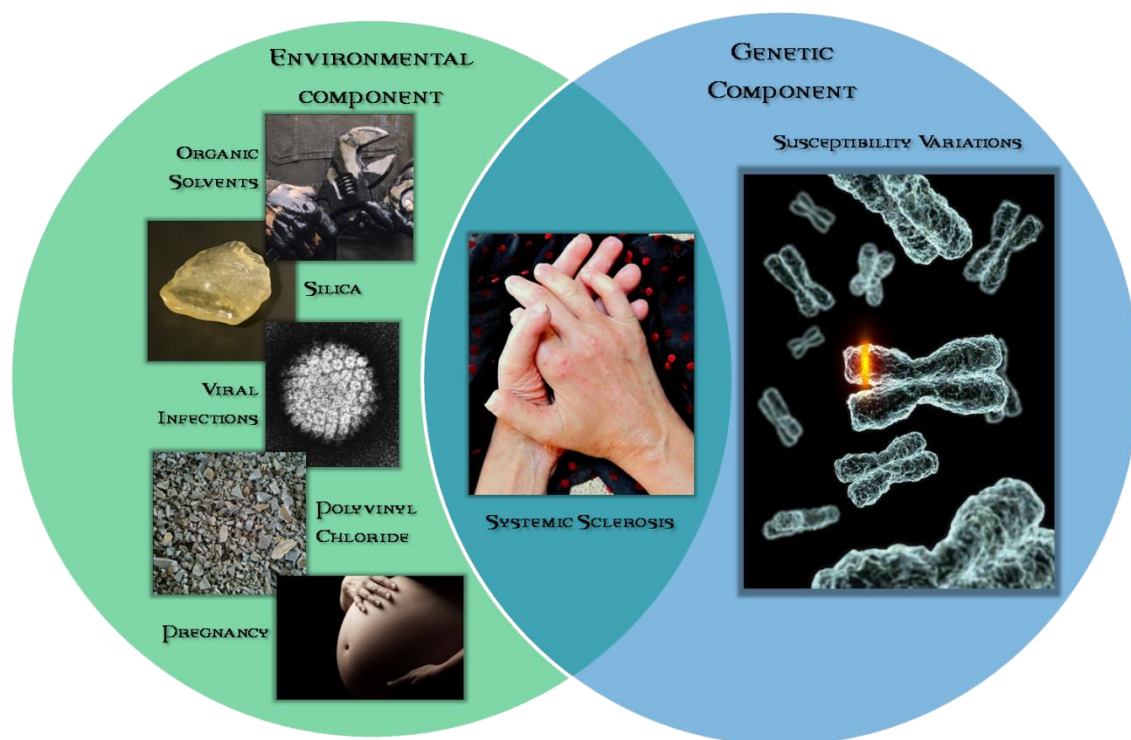


Figure 1. Factors which determine the development of SSc.

Thus, SSc is a complex clinically heterogeneous disease in which multiple genetic risk factors, each with a modest role in the disease risk, interact with environmental factors to trigger the onset of the disease and affect the severity and course of the disease (*figure 1*). Both, the genetic component and the environmental factors of the disease have been studied, but also both, especially the genetic component remains largely elusive.

As in most AIDs, a sex bias in SSc patients can be observed, affecting more women than men, typically in a 9:1 ratio as observed in the cohorts studied in this thesis [16-18] (*figure 2*). Although different mechanisms have been proposed for this sex proportion deviation in SSc patients [19-21], none have proven as solid evidence for this observation so far. A possible explanation for this increased proportion of women affected by SSc could be explained by pregnancy related pathological processes that has been observed in women with SSc such as fetal antimaternal graft-versus-host reactions [22] and the presence of DNA of the offspring in class II compatible women with their child [23]. Since those are pregnancy related processes, this could explain, at least partially, the sex bias present in patients with SSc.

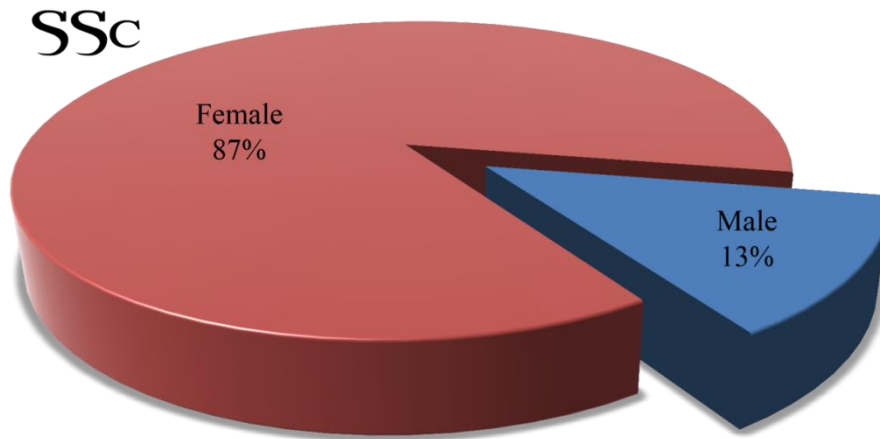


Figure 2. Gender distribution in the main SSc cohorts used in this thesis.

Figure 3 shows the distribution of the two main SSc subtypes in the case cohorts investigated during this thesis, while figure 4 shows the distribution of the two most frequent auto-antibodies (*i.e.* ACA and ATA) in this body of patients. How the clinical subtypes and the auto-antibody positive groups overlap is illustrated in figure 5.

All SSc patients in our cohorts either met the American College of Rheumatology Preliminary criteria for the classification of SSc or had at least three of the five CREST (calcinosis, Raynaud's phenomenon, esophageal dysmotility, sclerodactyly, telangiectasias) features [24]. Also, all individuals were of Caucasian origin, determined either by principal component analysis (for the individuals genotyped at GWAS level) or self-reported ancestry (for individuals of the replication cohorts).

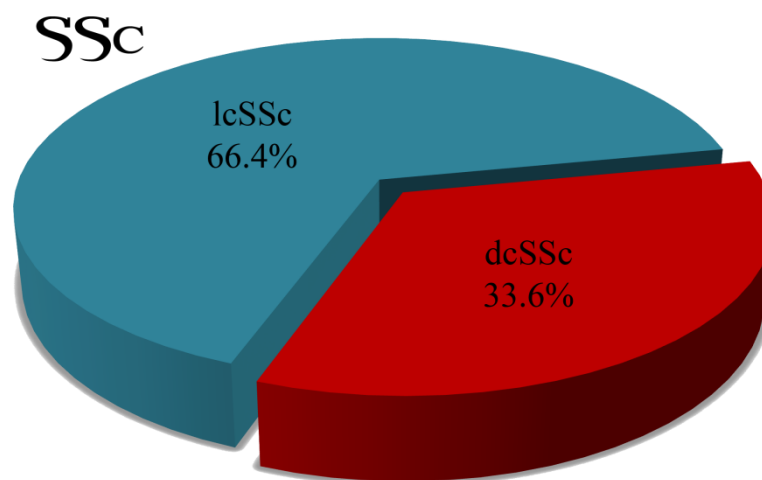


Figure 3. Subtype distribution in the main SSc cohorts used in this thesis.

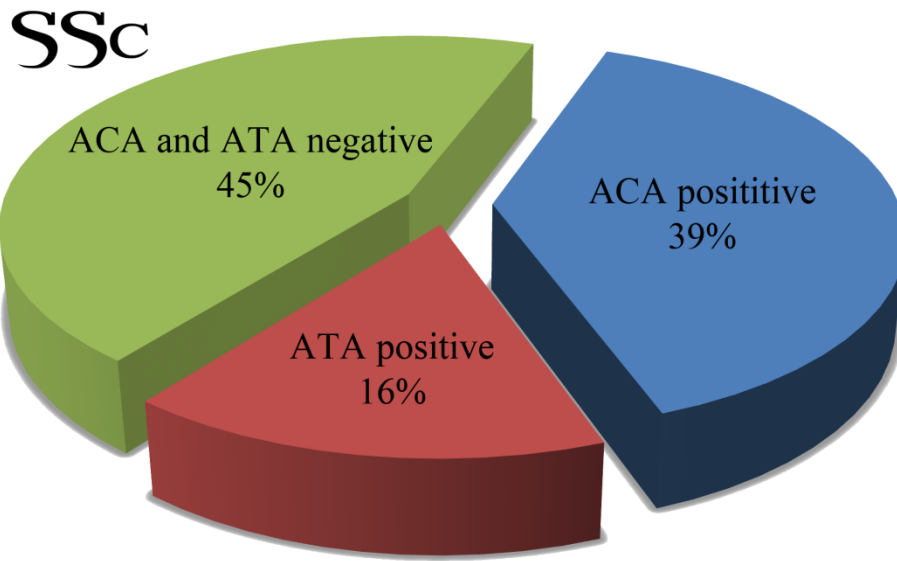


Figure 4. Auto-antibody distribution in the main SSc cohorts used in this thesis.

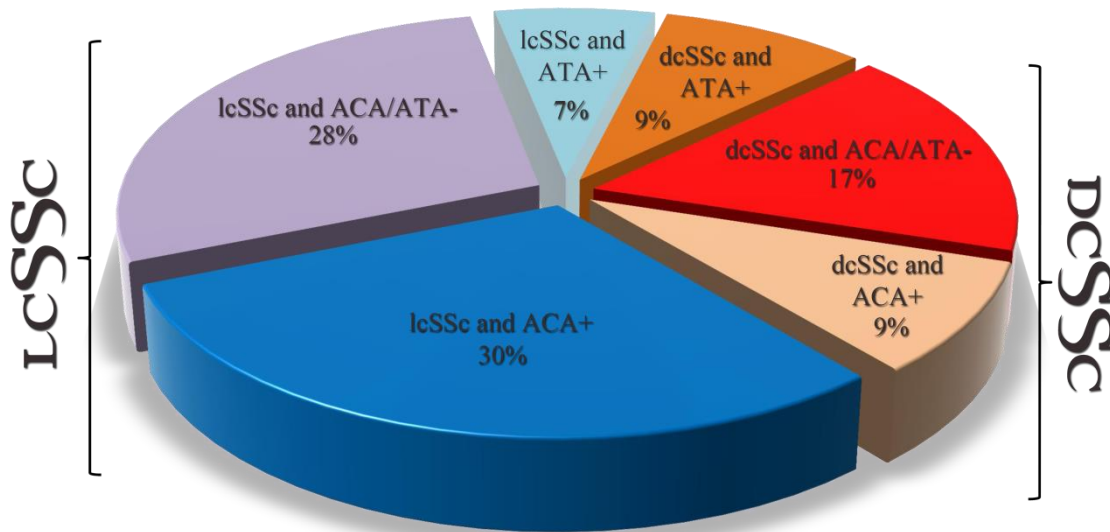


Figure 5. Subtype and auto-antibody distribution of the main cohorts used in this thesis. All percentages are relative to the total. All patients in our study cohorts were either classified as lcSSc or dcSSc. Of these, they could present ACA auto-antibodies, ATA auto-antibodies or none of them.

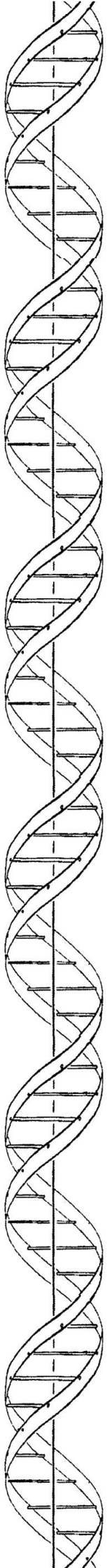
Environmental Factors in Systemic Sclerosis

As a complex trait, SSc is influenced by both genetic variation in individuals and their environment [25]. Several studies have been carried out in SSc in order to identify the environmental factors which affect the development of this disorder. In general, the loss of tolerance to self is the crucial factor which must occur in order to develop an AID such as SSc. The environment can affect the loss of tolerance in one of two ways: 1) alteration of self-antigens by substances of any kind, which makes the immune system recognize the modified molecular motif as alien or 2) a molecular mimicry process in which an environmental agent is cross-recognized with self-antigens.

Among the best described environmental factors which predispose to the development of SSc is the exposure to Silica dust [25, 26]. It has been demonstrated that human lymphocytes exposed to silica express high levels of *CD95* (the Fas receptor) which induces apoptosis along with different autoantigen alterations, which in turn provokes an autoimmune response [27] and post translational protein changes [28]. Studied environments under which SSc or SSc-like syndromes develop due to silica exposure include gold mines, powder factories, uranium mines and others [25]. Also the exposure to organic solvents such as benzene, toluene, xylene to name a few can cause similar phenotypes [25]. Additionally the exposure to the vinyl chloride monomer (found in the polyvinyl chloride dust) can cause an SSc-like syndrome [29-31]. These kind of exposures to different substances fall into the first category of environmental component of the alteration of the self-antigens which trick the immune system into recognizing them as alien.

When attending to the possibility of molecular mimicry as environmental factor influencing the development of SSc we find studies in which it has been described a sequence of 11 amino acids in the C-terminal end of the topoisomerase I (one of the two major auto-antibodies in SSc) with high homology with antigens of certain mammal retroviruses [32]. Another study was able to characterize regions of homology between the UL70 protein of human cytomegalovirus and other fragments of the human topoisomerase I [33].

As previously commented, pregnancy related pathological processes such as fetal antimaternal graft-versus-host reactions [22] and the presence of DNA of the offspring



in class II compatible women with their child [23] could also be included among the environmental factors affecting the onset of SSc.

An SSc-like syndrome epidemic was spread in Spain in 1981 due to the ingestion of oil denatured with 2% aniline. This epidemic, known as toxic oil syndrome, had female prevalence, different clinical evolution even inside the same family and HLA-DR2 was increased in patients, which points to an underlying genetic component for this syndrome as in the case of SSc [34, 35].

At last, unlike in RA, smoking has been posed as not influencing the development of SSc but affecting the severity of the disease, including parameters as Raynaud's phenomenon or pulmonary capacity in SSc patients [36, 37].

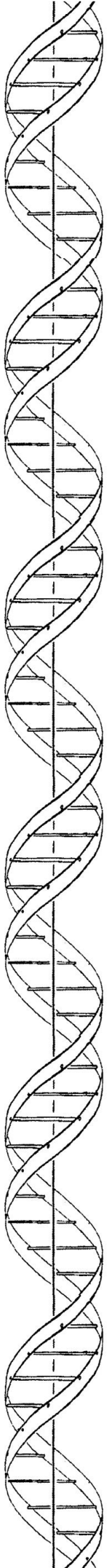
Thus the exposure to organic solvents, silica dust and polyvinyl chloride (through self-antigen modification), viral infections, pregnancy related microchimerism (through molecular mimicry) and others are environmental factors which influence the development of SSc. Nevertheless, it is striking that of them can explain the environmental component of SSc, albeit some of them can explain specific phenotypes presented in individuals suffering from SSc or SSc-like syndromes. Interestingly this mimics the trend of genetic variation explaining not the developing of the disease as a whole but more specific SSc phenotypes as we will discuss in this thesis.

Genetic Component of Systemic Sclerosis

Commonly, genetic susceptibility variants are searched for by obtaining the genotypes (with a plethora of methods available) of a certain amount of genetic variants in cohorts of individuals presenting the studied trait (cases) and individuals without it (controls). The frequency of the genetic variants of interest are then compared by the means of different statistics between cases and controls, and if those differences are significant according to the statistical test used, then it is assumed that the variant(s) are associated with a greater chance of presenting the trait. Statistical tests must be adjusted for different aberrations product of the sample size or the number of variants tested, *e.g.*, the test statistics must be corrected for multiple testing in order to avoid false positive signals as a consequence of the number of genetic variants tested.

As previously stated, SSc is a complex disorder of the immune system and the connective tissue with both a genetic and an environmental component [1]. Two facts point towards the weight of the genetic component in SSc: 1) the prevalence of the disease varies from 7 per million to 700 per million in different populations [18, 38-40] and 2) Twin and familiar studies revealed a high concordance of auto-antibody production and HLA-haplotypes, making the chance of affected siblings of developing SSc up to 15-fold [41-43].

The first genetic susceptibility locus which has been confirmed described for SSc were the HLA class II genes [12, 44], although not until recent more genes which affect the development and course of this disease have been discovered (*table 1*). Prior to the publication of the first SSc GWAS in a Caucasian population [45], few non-HLA susceptibility loci were involved in SSc (*table 1*). A GRAIL analysis [46] of this loci shows that most putatively responsible genes for the association observed in this regions belong to the immune system (*figure 4*). Of the three major hallmarks of SSc, collagen deposit, vascular damage and altered immune response [1], this can only partially explain the later one, leaving no genetic evidence as what genes directly influence the other two pathological processes in SSc, although an altered immune response can lead indirectly to vascular damage and fibrosis through inflammation. Furthermore, when merging the Gene Ontology (GO) terms of all the 16 loci associated with SSc prior to this thesis in a word cloud no single mention of fibrosis or collagen deposit is observed (*figure 5*).



Gene	Variation	Phenotype	OR	P value	References
<i>BANK1</i>	rs17266594	dcSSc	1.23	1.00x10 ⁻³	[47, 48]
<i>BLK</i>	rs2736340	ACA	1.47	2.20x10 ⁻⁶	[49-51]
HLA Class II	HLA-DRB1*1104	SSc	4.99	3.00x10 ⁻⁴	[12, 52, 53]
HLA Class II	HLA-DPB1*1301	ATA	14.02	<1.00x10 ⁻⁴	[12, 54]
HLA Class II	HLA-DQB1*0501	ACA	2.56	<1.00x10 ⁻⁴	[12, 53, 55]
<i>IL12RB2</i>	rs3790567	SSc	1.17	2.82x10 ⁻⁹	[56]
<i>IL2RA</i>	rs2104286	ACA	1.30	2.07x10 ⁻⁴	[57]
<i>IRAK1</i>	rs1059702	ATA	1.43	9.39x10 ⁻⁵	[58]
<i>IRF5</i>	rs10488631	SSc	1.50	1.86x10 ⁻¹³	[45, 59, 60]
<i>IRF7</i>	rs1131665	ACA	0.78	6.14x10 ⁻⁴	[61]
<i>STAT4</i>	rs3821236	SSc	1.30	3.37x10 ⁻⁹	[62-64]
<i>TNFAIP3</i>	rs5029939	dcSSc	1.46	2.29x10 ⁻⁶	[65, 66]
<i>TNFSF4</i>	rs12039904	ACA	1.22	2.09x10 ⁻³	[67, 68]
<i>TNIP1</i>	rs4958881	ATA	1.19	3.26x10 ⁻⁵	[69, 70]

Table 1. Genetic loci associated with susceptibility to SSc or its considered subphenotypes prior to the first SSc GWAS in a Caucasian population. The genes shown are selected from each region by GRAIL analysis. The considered cutoff P value was 5x10⁻³.

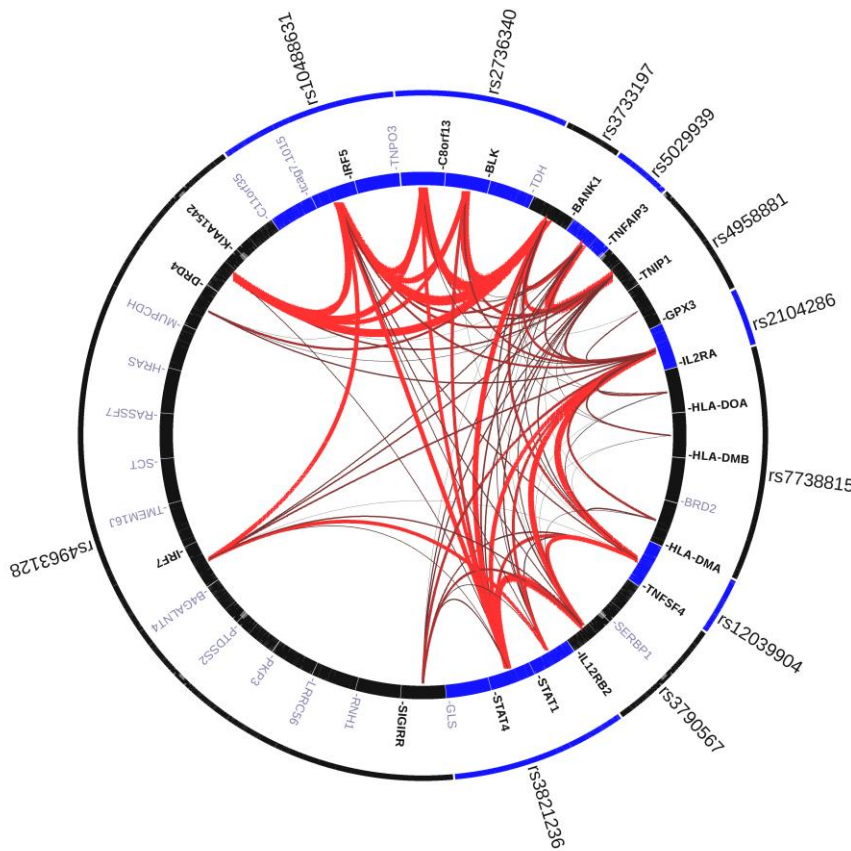


Figure 6. GRAIL analysis of the genetic loci associated with SSc or any of its considered subphenotypes prior to the realization of the first GWAS in a Caucasian population. The release 18 of the human genome and the PubMed text as of 2012 were used.

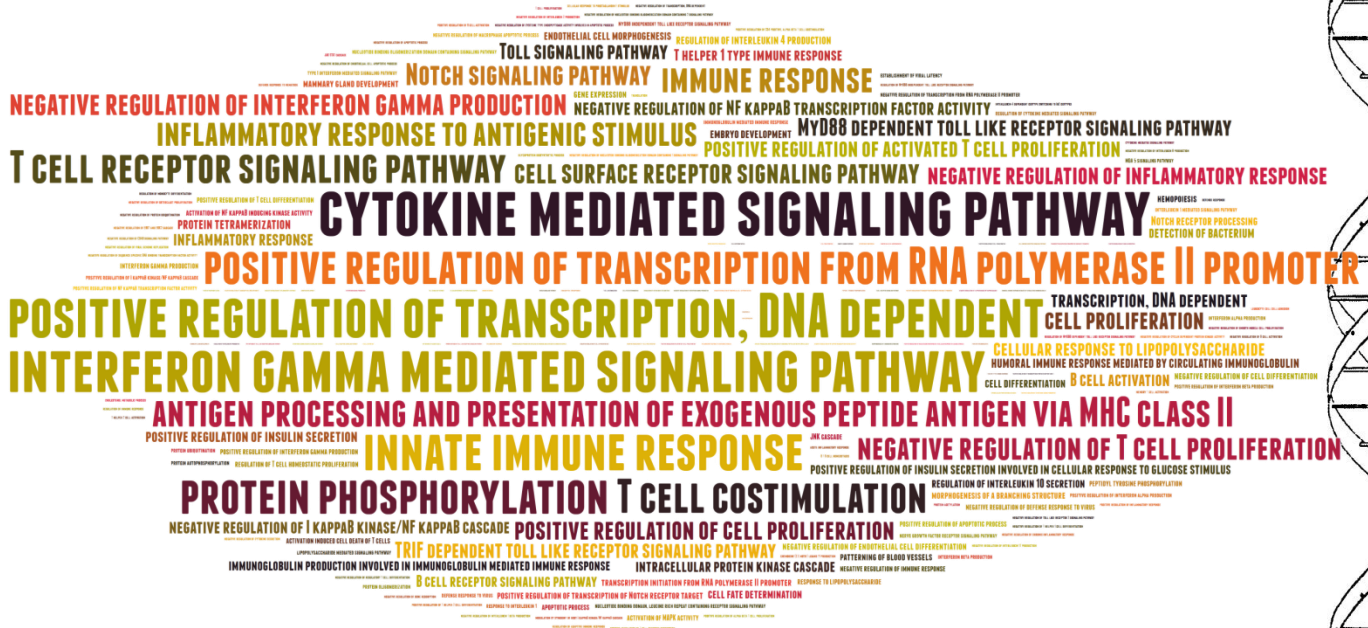


Figure 7. Word cloud representing the GO terms of all the genes associated with SSc or its considered subphenotypes prior to the first SSc GWAS in a Caucasian population according to GRAIL selection. The size of each GO term is weighted according to the number of occurrences in all GO terms from the 16 loci.

SSc is a heterogeneous disorder, which main clinical division are the limited cutaneous subtype (lcSSc) and the diffuse cutaneous subtype (dcSSc) [1]. Furthermore, the disease presents two major auto-antibodies, which are associated to disease outcome: ACA and ATA [1]. Several of the described SSc genetic associations are confined to one of these subgroups within the disease, adding further complexity to the genetic component of SSc (table 1 and figure 6). The fact that most of the described genetic associations within these subphenotypes belong to lcSSc and less to the ATA producing patients does not point to a weaker genetic component in these compartments, but to a lesser genetic power in the smaller subgroups, adding additional difficulty to detect clear signals of association. Nevertheless, the most important associations described in SSc, those of the HLA class II alleles, have been mostly describe to influence not the overall disease but the auto-antibody positive subgroups (figure 6) [12, 52, 55]. This points in an interesting direction: SSc, lcSSc and dcSSc are clinical entities defined by clinicians due to the necessity in medical practice to classify the patients into diseases or disorders, but when attending to biological processes (such as auto-antibody production), more homogeneous groups are observed, genetically speaking.

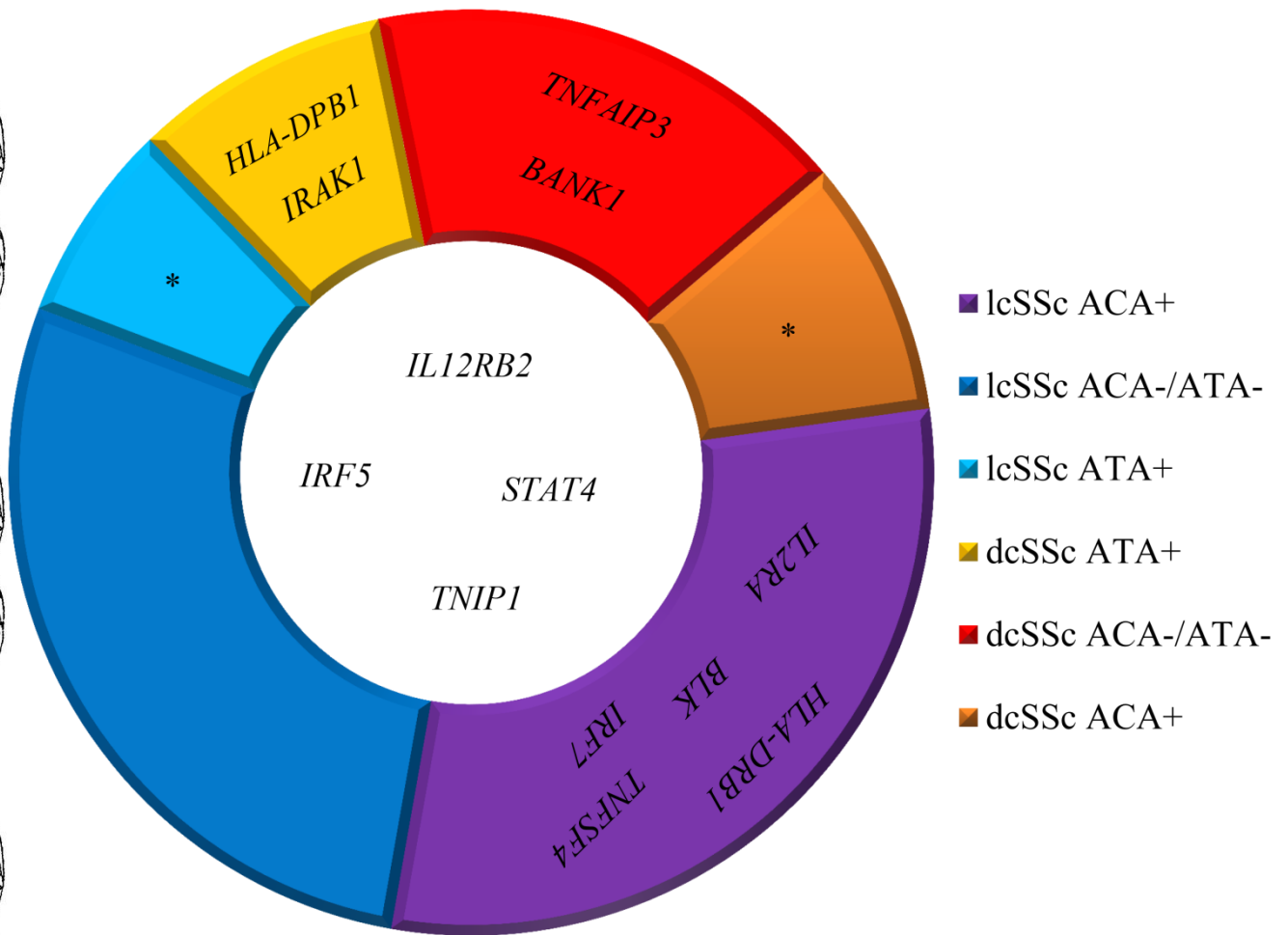


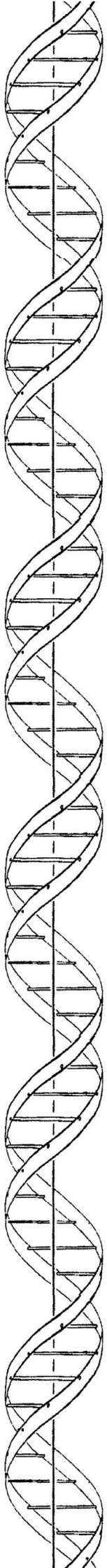
Figure 8. Diagram showing the genetic loci associated with SSc or its considered subphenotypes prior to the first SSc GWAS in a Caucasian population divided according disease subtype and/or auto-antibody production. *The lcSSc/ATA+ and dcSSc/ACA+ are not traditionally analyzed, because although this combinations of subtype and auto-antibody do exist, lcSSc is more commonly accompanied by ACA and dcSSc is more commonly accompanied by ATA; thus, the ACA and ATA specific associations (depicted here in the lcSSc/ACA and dcSSc/ATA) should correspond also to this segment.

THE ADVENT OF THE GENOMIC ERA

In a Genome-Wide Association Study (GWAS) hundreds of thousands or even millions of variants are genotyped through different technologies in large case-control cohorts. This allows us to interrogate the whole genome in a hypothesis free fashion in order to uncover genetic susceptibility variants influencing any given trait.

Since the publication of the first GWAS in 2005 by Klein *et al.* in macular degeneration [71], an avalanche of GWAS have been performed in multitude of human normal traits, such as, color of the eyes, hair and skin, fat distribution and height [72-74]; and human disorders such as autoimmune disorders, cardiovascular disorders, bipolar disorders and cancer [75-81]. Many new genes, implicating new metabolic pathways and physiological processes have been involved in the pathogenesis of several complex human diseases such as rheumatoid arthritis or systemic lupus erythematosus [78, 80]. Nevertheless, the scientific community is far from fulfilling the promise of the complete understanding of the genetic mechanisms underlying the onset of such diseases [82]. The genetic component of these human traits has been far more elusive than anticipated, mainly because of 1) the GWAS platforms used to date do not account for rarer genetic variations which now are believed to play an important role in the pathogenesis of human complex diseases, 2) the lack of new reliable and reproducible statistical and bioinformatical methods with which to analyze the new kind of data, 3) the lack of powerful enough computers to properly analyze the dramatically increasing amount of genetic data, 4) the fact that many common genetic variants remain to be included in genotyping platforms as the 1,000 genomes project is unveiling, and 5) the scarce phenotyping available when analyzing the genomic data [82, 83].

When performing a candidate gene study, or even a fine mapping study centered on a gene or a genetic region, typically from one to several hundreds of genetic variants are genotyped and analyzed. This relatively low number of analyzed variants does not generate a compromising level of false positives. However, in a GWAS the genotyping of hundreds of thousands to millions of variants are generated and analyzed, and the multiplicity of tests will provoke many false positive significant signals in the analysis. To partially solve this, strict levels of correction based on the number of tests performed must be applied to correct GWAS analyses. Most typical GWAS threshold for



significance is based on the number of independent LD blocks present in the human genome (which are thus considered independent tests), which gives us a threshold for the significance of the P values fixed in 5×10^{-8} , instead of the traditional 0.05. This significance level has been largely known as GWAS significance level.

In the first RA GWAS performed, only the HLA region, *PTPN22* and the 6q23 region showed a GWAS significant level signal [75]. However, more than 35 genetic regions are known to influence RA risk now [84, 85], being those association signals in the GWAS study in what we will call the grey zone of association. P values between 5×10^{-8} and 0.05 are mostly false positive signals, but many of the numerous relatively low risk genetic variants influencing disease susceptibility are found in that range of association in GWAS data. Several approaches have been used in order to extract the RA genetic risk factors from this grey zone. Firstly, and the most obvious, is to increase the statistical power of the GWAS: with this approach *REL* was determined as an RA susceptibility gene by expanding the existent North America Rheumatoid Arthritis Consortium (NARAC) [80]. Secondly, the selection of a reasonable amount of SNPs from the grey zone under different criteria has been fruitful in identifying *MSRA* (pathway enrichment and replication), *CD28*, *PRDM1* and *CD2/CD58* (GRAIL enrichment and replication) as RA susceptibility genes [86, 87]. Thirdly, the analysis of a RA subphenotype (the presence of Anti Citrullinated Protein Auto-antibody, or ACPA) has also provided invaluable insight of the genetics of this disease [88-90].

In SSc genetics there has been an explosion of the number of genetic susceptibility loci in the last years, since the publication of the first GWAS in 2010 (*figure 7*) [45]. This number has passed from a few (*table 1*) to 26 including the susceptibility loci described in this thesis. This has been possible due to an international effort of several groups which together have reunited large cohorts from different countries, with enough statistical power to design studies in which the genetic component of SSc is slowly but firmly being revealed. Even then, the promise of the genomic era and the GWAS has not been fulfilled, and much remains to be uncovered in the genetics of SSc and human complex traits.

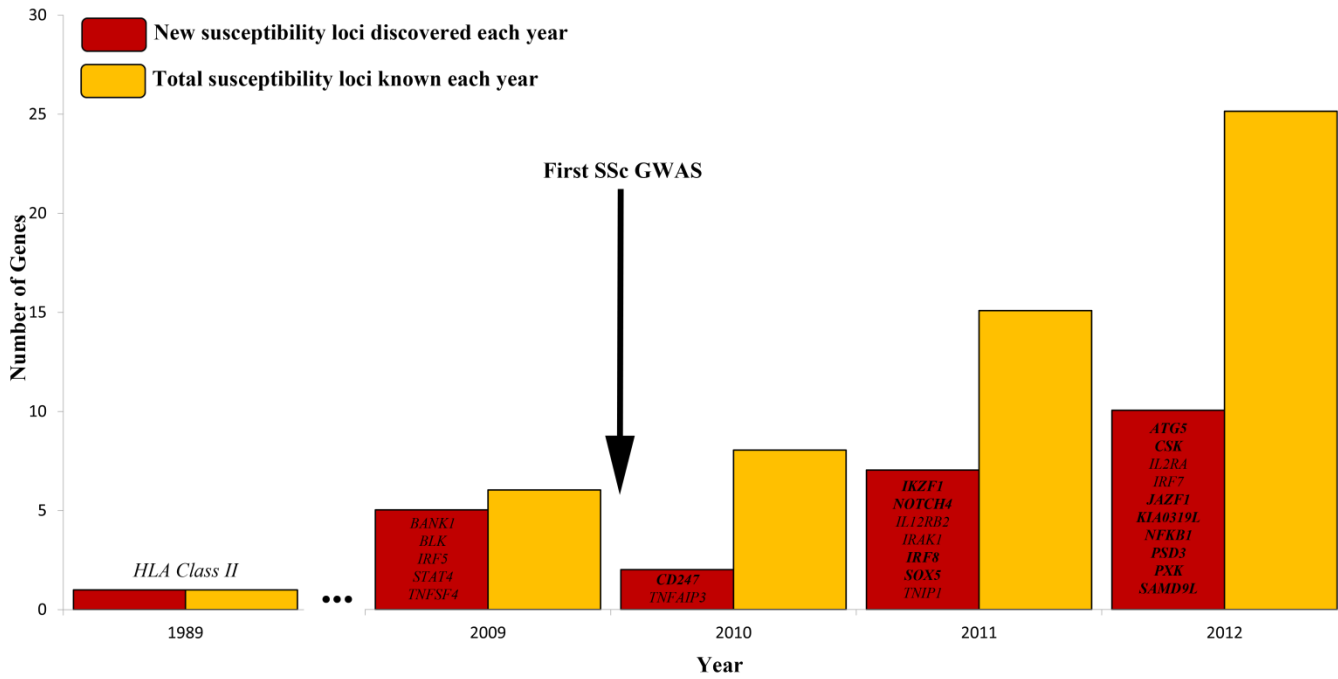
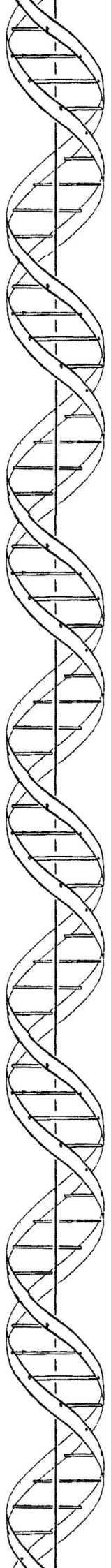


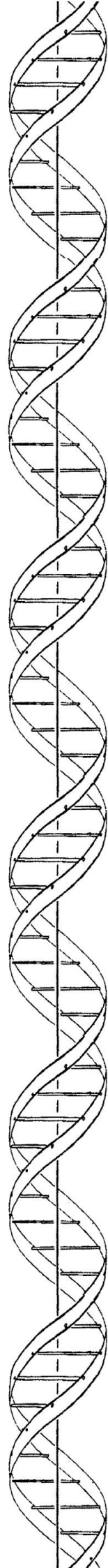
Figure 9. Bar plot showing the number of new susceptibility loci discovered each year (in red) and the total susceptibility loci known each year (in yellow) for SSc or any of its considered subphenotypes. Only the genes from studies with more than 1,000 SSc, replicated in more than one population and with a P value lower than 5×10^{-3} are considered as established susceptibility loci, and thus, are shown in this figure. In **Bold** are marked the novel susceptibility loci described in this thesis.



I CAN ACCEPT FAILURE, BUT I CAN'T ACCEPT NOT TRYING.

-MICHAEL JORDAN

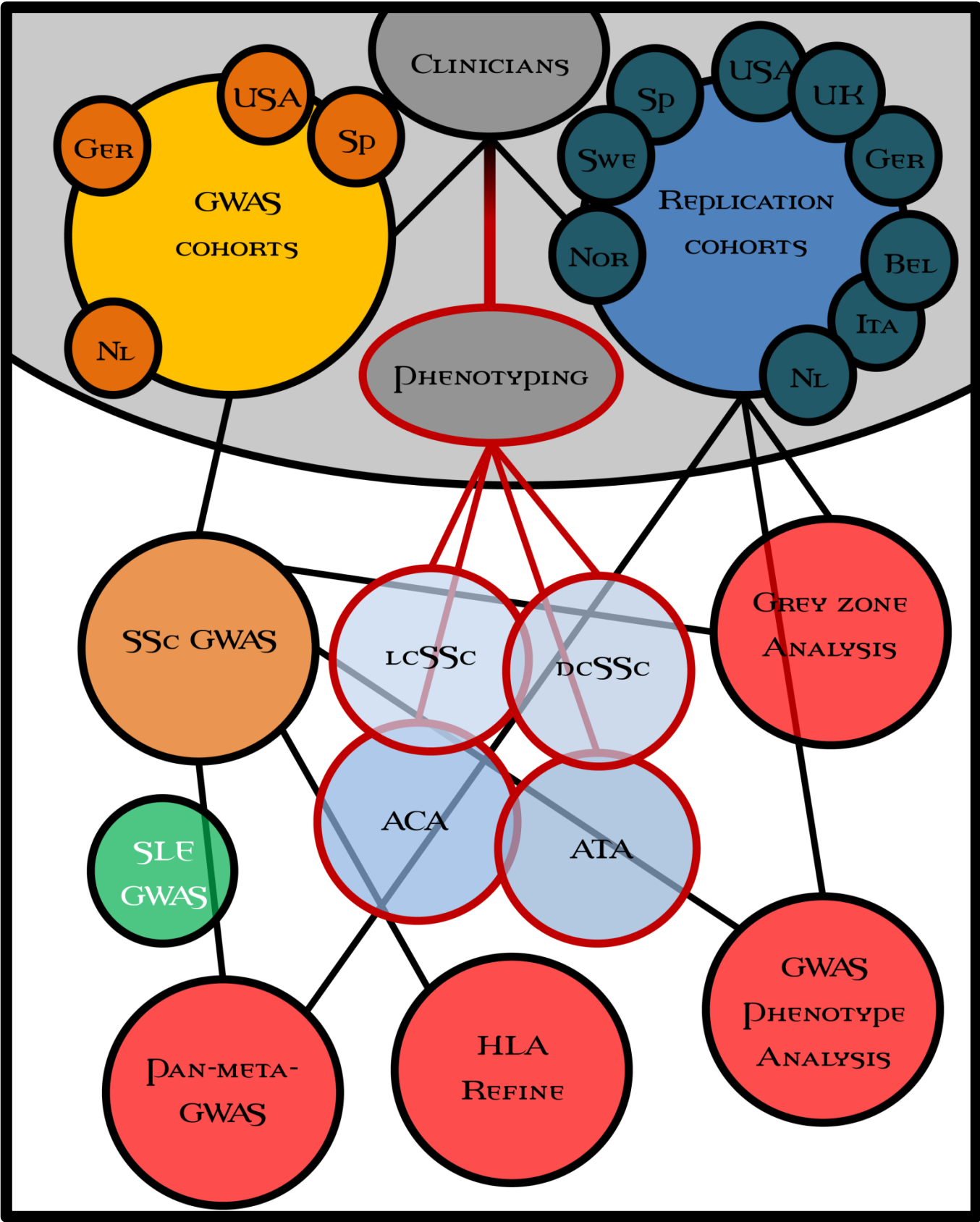
OBJECTIVES

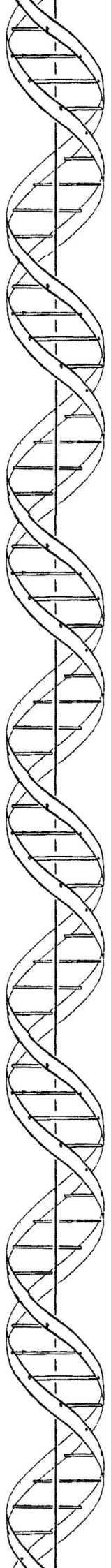


OBJECTIVES

TO DETERMINE THE **GENETIC COMPONENT OF SYSTEMIC SCLEROSIS** AND ITS MAJOR SUBPHENOTYPES, MORE SPECIFICALLY:

1. TO FIND NEW **GENETIC SUSCEPTIBILITY VARIANTS** WHICH INFLUENCE THE DEVELOPMENT SYSTEMIC OF SCLEROSIS.
2. TO FIND NEW SUSCEPTIBILITY GENETIC VARIANTS WHICH DIFFERENTIATE SYSTEMIC SCLEROSIS MAJOR **SUBPHENOTYPES**, *I.E.* ANTI-TOPOISOMERASE I AUTO-ANTIBODIES, ANTI-CENTROMERE AUTO-ANTIBODIES, LIMITED CUTANEOUS SUBTYPE AND DIFFUSE CUTANEOUS SUBTYPE.
3. TO EXPLORE THE **GREY ZONE OF ASSOCIATION** IN THE FIRST SSc GWAS AND EXTRACT THE TRUE SUSCEPTIBILITY GENETIC VARIANTS THEREIN.
4. TO REFINE THE LARGELY KNOWN **HLA ASSOCIATIONS** WITH SYSTEMIC SCLEROSIS DOWN TO THE AMINOACID LEVEL AND DETERMINE WHETHER THEY BELONG INTO THE TOTAL SSc OR ANY OF ITS SUBPHENOTYPES.
5. TO FIND HOW THE SUSCEPTIBILITY GENES ARE **SHARED AMONG AUTOIMMUNE DISEASES**.

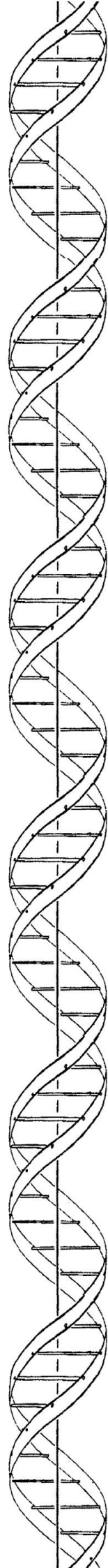




THE PLURAL OF ANECDOTE IS NOT DATA.

-ROGER BRINNER

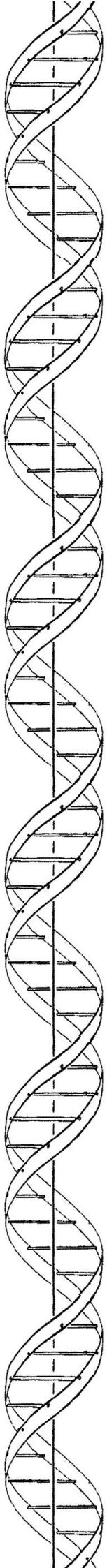
PUBLICATIONS

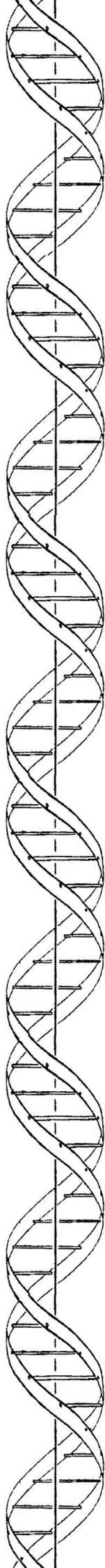


PUBLICATIONS



GENOME-WIDE ASSOCIATION STUDY OF SYSTEMIC
 SCLEROSIS IDENTIFIES *CD247* AS A NEW
 SUSCEPTIBILITY LOCUS. *NATURE GENETICS*, 2010.





Genome-wide association study of systemic sclerosis identifies *CD247* as a new susceptibility locus

Timothy R D J Radstake^{1,38}, Olga Gorlova^{2,38}, Blanca Rueda^{3,38}, Jose-Ezequiel Martin^{3,38}, Behrooz Z Alizadeh⁴, Rogelio Palomino-Morales⁵, Marieke J Coenen⁵, Madelon C Vonk¹, Alexandre E Voskuyl⁶, Annemie J Schuerwegh⁷, Jasper C Broen¹, Piet L C M van Riel¹, Ruben van 't Slot⁴, Annet Italiaander⁴, Roel A Ophoff^{4,8}, Gabriela Riemekasten⁹, Nico Hunzelmann¹⁰, Carmen P Simeon¹¹, Norberto Ortego-Centeno¹², Miguel A González-Gay¹³, María F González-Escribano¹⁴, Spanish Scleroderma Group³⁷, Paolo Airo¹⁵, Jaap van Laar¹⁶, Ariane Herrick¹⁷, Jane Worthington¹⁷, Roger Hesselstrand¹⁸, Vanessa Smith¹⁹, Filip de Keyser¹⁹, Fredric Houssiau²⁰, Meng May Chee²¹, Rajan Madhok²¹, Paul Shiels²¹, Rene Westhovens²², Alexander Kreuter²³, Hans Kiener²⁴, Elfride de Baere²⁵, Torsten Witte²⁶, Leonid Padykov²⁷, Lars Klareskog²⁷, Lorenzo Beretta²⁸, Raffaella Scorza²⁸, Benedicte A Lie²⁹, Anna-Maria Hoffmann-Vold³⁰, Patricia Carreira^{13,31}, John Varga³², Monique Hinchcliff³², Peter K Gregersen³³, Annette T Lee³³, Jun Ying², Younghun Han², Shih-Feng Weng², Christopher I Amos², Fredrick M Wigley³⁴, Laura Hummers³⁴, J Lee Nelson³⁵, Sandeep K Agarwal³⁶, Shervin Assassi³⁶, Pravitt Gourh³⁶, Filemon K Tan³⁶, Bobby P C Koeleman^{4,38}, Frank C Arnett^{36,38}, Javier Martin^{3,38} & Maureen D Mayes^{36,38}

Systemic sclerosis (SSc) is an autoimmune disease characterized by fibrosis of the skin and internal organs that leads to profound disability and premature death. To identify new SSc susceptibility loci, we conducted the first genome-wide association study in a population of European ancestry including a total of 2,296 individuals with SSc and 5,171 controls. Analysis of 279,621 autosomal SNPs followed by replication testing in an independent case-control set of European ancestry (2,753 individuals with SSc (cases) and 4,569 controls) identified a new susceptibility locus for systemic sclerosis at *CD247* (1q22–23, rs2056626, $P = 2.09 \times 10^{-7}$ in the discovery samples, $P = 3.39 \times 10^{-9}$ in the combined analysis). Additionally, we confirm and firmly establish the role of the MHC ($P = 2.31 \times 10^{-18}$), *IRF5* ($P = 1.86 \times 10^{-13}$) and *STAT4* ($P = 3.37 \times 10^{-9}$) gene regions as SSc genetic risk factors.

SSc is a profoundly disabling autoimmune disease characterized by vascular damage, altered immune responses and abnormal fibrosis of skin and internal organs leading to premature death in affected individuals¹. The etiology of SSc is complex and poorly understood, but as with most autoimmune conditions, it is widely accepted that environmental genetic factors contribute to disease risk. Data from familial, twin and ethnicity studies support the relevance of the genetic component in SSc etiology². Previous studies aimed at dissecting the genetic factors underlying SSc genetic susceptibility have used the candidate gene association study approach³. In spite of several years of research, this strategy has yielded a very limited characterization of SSc genetic risk factors. Except for the major histocompatibility complex (MHC) genes, which are relevant genetic markers for SSc across populations, few other loci outside the human leukocyte antigen (HLA) region have demonstrated strong and reproducible associations with SSc

¹Department of Rheumatology, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands. ²Department of Epidemiology, M.D. Anderson Cancer Center, Houston, Texas, USA. ³Instituto de Parasitología y Biomedicina López-Neyra, Consejo Superior de Investigaciones Científicas, Granada, Spain. ⁴Department of Medical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands. ⁵Department of Human Genetics, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands. ⁶Department of Rheumatology, Vrije Universiteit (VU) Medical Centre, Nijmegen, The Netherlands. ⁷Department of Rheumatology, University of Leiden, Nijmegen, The Netherlands. ⁸University of California Los Angeles Center for Neurobehavioral Genetics, Los Angeles, California. ⁹Department of Rheumatology and Clinical Immunology, Charité University Hospital, Berlin, Germany. ¹⁰Department of Dermatology, University of Cologne, Cologne, Germany. ¹¹Servicio de Medicina Interna, Hospital Valle de Hebrón, Barcelona, Spain. ¹²Servicio de Medicina Interna, Hospital Clínico Universitario, Granada, Spain. ¹³Servicio de Reumatología, Hospital Marqués de Valdecilla, Santander, Spain. ¹⁴Servicio de Inmunología, Hospital Virgen del Rocío, Sevilla, Spain. ¹⁵University of Brescia, Brescia, Italy. ¹⁶University of Newcastle, Newcastle Upon Tyne, UK. ¹⁷Department of Rheumatology and Epidemiology, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK. ¹⁸University of Lund, Lund, Sweden. ¹⁹University of Ghent, Ghent, Belgium. ²⁰University of Leuven, Leuven, Belgium. ²¹University of Glasgow, Glasgow, UK. ²²University of Antwerpen, Antwerpen, Belgium. ²³Ruhr University of Bochum, Bochum, Germany. ²⁴University of Vienna, Vienna, Austria. ²⁵Department of Genetics, University of Ghent, Ghent, Belgium. ²⁶University of Hannover, Hannover, Germany. ²⁷Karolinska Institute, Stockholm, Sweden. ²⁸University of Milan, Milan, Italy. ²⁹Institute of Immunology and ³⁰Department of Rheumatology, Rikshospitalet, Oslo University Hospital, Oslo, Norway. ³¹Hospital 12 de Octubre, Madrid, Spain. ³²Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA. ³³Feinstein Institute of Medical Research, Manhasset, New York, USA. ³⁴Johns Hopkins University Medical Center, Baltimore, Maryland, USA. ³⁵Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. ³⁶The University of Texas Health Science Center–Houston, Houston, Texas, USA. ³⁷A full list of members is provided in the **Supplementary Note**. ³⁸These authors contributed equally to this work. Correspondence should be addressed to T.R.D.J.R. (tradstake73@gmail.com) or M.D.M. (maureen.d.mayes@uth.tmc.edu).

Received 4 January; accepted 4 March; published online 11 April 2010; corrected online 16 April 2010; doi:10.1038/ng.565

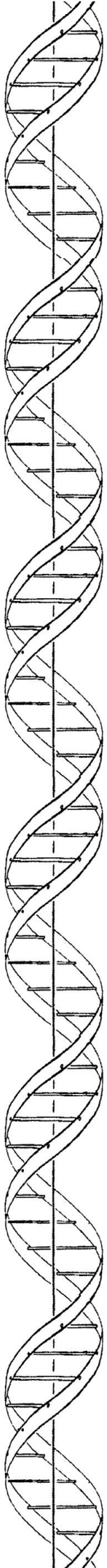


Table 1 Loci showing the strongest association signal with SSc susceptibility outside the MHC region

Chr.	Gene	SNP	Location	Position	Minor allele	MAF (case/control)	GC-corrected P value	PC-corrected P value	OR (95% CI)
7q32	<i>TNPO3-IRF5</i>	rs10488631	Downstream	128,381,419	C	0.145/1.102	1.86×10^{-13}	3.84×10^{-14}	1.50 (1.35–1.67)
		rs12537284	Intergenic	128,505,142	A	0.162/0.129	2.74×10^{-7}	1.49×10^{-7}	1.30 (1.18–1.44)
		rs4728142	Upstream	128,361,203	A	0.494/0.445	5.21×10^{-7}	1.81×10^{-7}	1.21 (1.12–1.29)
2q32	<i>STAT4</i>	rs3821236	Intronic	191,611,003	A	0.247/0.202	3.37×10^{-9}	3.93×10^{-9}	1.30 (1.19–1.41)
1q22–23	<i>CD247</i>	rs2056626	Intronic	165,687,049	G	0.370/0.421	2.09×10^{-7}	3.27×10^{-7}	0.82 (0.76–0.88)
18q22	<i>CDH7</i>	rs10515998	Intergenic	61,521,202	G	0.062/0.040	2.25×10^{-7}	1.01×10^{-7}	1.53 (1.31–1.79)
6p25	<i>EXOC2-IRF4</i>	rs4959270	Intronic	402,748	A	0.445/0.494	1.23×10^{-7}	9.06×10^{-8}	0.82 (0.77–0.88)

Chr., chromosome; BP, base pairs; MAF, minor allele frequency; GC, genomic control; PC, principal component; OR, odds ratio.

susceptibility^{3,4}. Only very recently have large case-control association studies identified *STAT4* and *IRF5* as genetic factors contributing to SSc susceptibility^{5–8}. As for other complex genetic disorders, it is expected that several genetic markers contribute to SSc predisposition with modest effects and therefore large sample sizes are required to detect new disease-associated loci⁹.

Therefore, we aimed more comprehensively to identify new SSc susceptibility loci and thus conducted the first genome-wide association study (GWAS) of SSc, including a total of 2,296 SSc cases and 5,171 healthy controls from four case-control series of European ancestry (from United States, Spain, Germany and The Netherlands) (Supplementary Table 1). Genotyping of the SSc case sets and Spanish controls was performed using the Illumina Bead-Array platform with chips of different SNP densities (Supplementary Table 1). The genotypes of the US controls were obtained from the Cancer Genetic Markers of Susceptibility (CGEMS) studies and the Illumina iControlDB database; German and Dutch control groups were extracted from previous studies or public databases^{10–13}.

After rigorous genotyping quality-control filters, a total of 279,621 SNPs shared between the four case-control series were extracted for analysis (Supplementary Table 1).

The genomic inflation factor (λ) was estimated for the complete combined dataset and showed evidence of a modest inflation of test statistics ($\lambda = 1.069$). When the HLA region was excluded from the analysis, the inflation of test statistics somewhat decreased ($\lambda = 1.066$) (Supplementary Fig. 1). To adjust for potential population stratification, we applied a genomic control correction to the test statistics. The potential effect of population substructure was tested by deriving principal components on a population-specific basis. We observed that case and control individuals in each population were not significantly different on the basis of these principal components and were therefore well genetically matched. We also performed an inverse variance-based meta-analysis, adjusting the odds ratios for the first five country-specific principal components. This analysis showed little variation from genomic control-corrected *P* values (Table 1).

The Mantel-Haenszel test under an allelic model revealed several SNPs reaching *P* values at genome-wide significance after genomic-control correction ($P \leq 5 \times 10^{-7}$) (Fig. 1). The strongest association signal was observed for a cluster of SNPs in an extended region at the 6p21 locus within the MHC region, whereas the rs6457617 SNP located in the *HLA-DQB1* gene region gave the highest *P* value (*P* genomic control-corrected = 2.31×10^{-18}) (Fig. 1 and Supplementary Table 2). Outside the MHC region, five loci showed association at $P < 10^{-7}$, namely the *TNPO3-IRF5* region in 7q32, *STAT4* in 2q32, *CD247* in 1q22–23, *CDH7* in 18q22 and *EXOC2-IRF4* near 6p25. The trend observed for all these loci was consistent across the different study populations (Supplementary Table 3). Furthermore, the *TNPO3-IRF5* locus obtained genome wide significance in the single US cohort and was further corroborated in the European cohorts (Supplementary Table 3).

SNPs mapping to the region of *TNPO3-IRF5* and *STAT4* achieved the strongest association observed for non-HLA genes (rs10488631, $P = 1.86 \times 10^{-13}$, OR = 1.50, 95% CI 1.35–1.67; rs3821236, $P = 3.37 \times 10^{-9}$, OR = 1.30, 95% CI 1.18–1.44) (Table 1 and Supplementary Table 3). Therefore, these results confirm the previously reported role of the MHC region, *STAT4* and *IRF5* as genetic risk factors for SSc and identify three new candidate loci^{3–8}.

We next aimed to confirm the association of the *CD247*, *CDH7* and *EXOC2-IRF4* loci with SSc susceptibility using a large independent replication case-control set comprising 2,753 SSc cases and 4,569 controls of European ancestry (Supplementary Table 4). The SNPs showing the strongest GWAS association on each region (rs2056626 for *CD247*, rs10515998 for *CDH7* and rs4959270 for *EXOC2-IRF4*) were genotyped in the replication cohorts using TaqMan 5' allelic discrimination assay technology. The association analysis by a Mantel-Haenszel test revealed a significant association of the rs2056626 genetic variant in the *CD247* region ($P = 3.07 \times 10^{-3}$, OR = 0.89, 95% CI 0.83–0.96) (Table 2 and Fig. 2). The combined analysis of the GWAS and replication cohort for this SNP revealed a highly significant association ($P = 3.39 \times 10^{-9}$, OR = 0.86, 95% CI 0.81–0.90). The association of the SNPs in the *CDH7* and *EXOC2-IRF4* regions was not confirmed in this replication cohort (Table 2 and Supplementary Fig. 2). Considering that the frequency observed for the *CDH7* rs10515998 genetic variant is quite low (around 5%), the population size of the replication cohort reached only 13% statistical power to detect an association at a significance level similar to that observed in the replication analysis (OR = 1.05). Therefore, the possible implication of the *CDH7* locus in SSc genetic predisposition

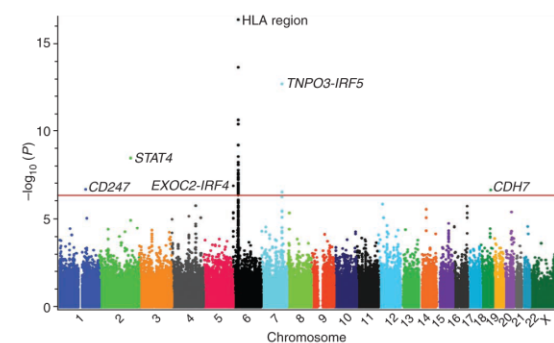


Figure 1 Manhattan plot of the GWAS of the discovery cohort comprising 2,346 SSc cases and 5,193 healthy controls. The $-\log_{10}$ of the Mantel-Haenszel test *P* value of 279,621 SNPs after correction by λ is plotted against its physical chromosomal position. Chromosomes are shown in alternate colors. SNPs above the red line represent those with a *P* value $< 5 \times 10^{-7}$. Plot corresponds to the combined analysis of the study cohorts.

LETTERS

Table 2 Association results for three loci genotyped in the replication samples

Chr.	Gene	SNP	Position	Minor allele	Stage	<i>n</i> (case/control)	MAF (case/control)	<i>P</i> value	OR (95% CI)
1q22–23	CD247	rs2056626	165,687,049	G	GWAS	2,296/5,014	0.370/0.421	2.09×10^{-7}	0.82 (0.76–0.88)
					Replication	2,566/4,387	0.366/0.394	3.07×10^{-3}	0.89 (0.83–0.96)
					Combined	4,867/9,401	0.368/0.409	3.39×10^{-9}	0.86 (0.81–0.90)
18q22	CDH7	rs10515998	61,521,202	G	GWAS	2,296/5,014	0.062/0.040	2.25×10^{-7}	1.53 (1.31–1.79)
					Replication	2,594/4,414	0.058/0.056	4.98×10^{-1}	1.05 (0.91–1.22)
					Combined	4,895/9,428	0.060/0.048	3.99×10^{-5}	1.25 (1.13–1.40)
6p25	EXOC2/IRF4	rs4959270	402,748	A	GWAS	2,296/5,171	0.445/0.494	1.23×10^{-7}	0.82 (0.77–0.88)
					Replication	2,361/4,372	0.466/0.469	6.34×10^{-1}	0.98 (0.91–1.05)
					Combined	4,662/9,554	0.456/0.483	2.16×10^{-5}	0.90 (0.85–0.94)

Chr., chromosome; MAF, minor allele frequency; OR, odds ratio.

should be further investigated. In contrast, because of the high minor allele frequency (MAF) of the rs4959270 polymorphism in the *EXOC2-IRF4* region, great heterogeneity of the association was observed in the replication cohorts (Supplementary Fig. 1). These findings are concordant with previous GWAS studies in which great population allelic heterogeneity has been reported for *EXOC2-IRF4* genetic variants leading to false positive disease associations, as may have occurred in our screening phase¹⁴. Notably, the newly identified SSc susceptibility locus, *CD247*, encodes a protein that participates in the regulation of immune response and thus could have a role in SSc pathogenesis. *CD247* encodes the T-cell receptor zeta (CD3 ζ) subunit, a component of the T-cell receptor (TCR)-CD3 complex¹⁵. The CD3 ζ chain plays an important role in the assembly of the TCR-CD3 complex and its transport to the cell surface and is crucial to receptor signaling function. It has been observed that the expression of the CD3 ζ chain is altered in chronic autoimmune and inflammatory disorders and that its low expression results in impaired immune response^{16–18}. Notably, *CD247* has been associated with susceptibility to systemic lupus erythematosus, another systemic autoimmune disease^{19,20}. Moreover, genetic variants in the 3' untranslated region of this gene have shown functional implications leading to a reduced expression of the CD3 ζ chain that could be manifested in systemic autoimmunity¹⁹. Therefore, further studies aiming to dissect the exact role of this molecule in SSc will be of interest.

This work represents the first large GWAS study conducted to date in SSc. Of note, the results obtained confirm and firmly establish the role of the HLA region, *STAT4* and *IRF5* in the genetic predisposition to SSc; these loci are also known to be risk factors for several other autoimmune conditions. In addition, a new susceptibility locus not previously

considered as a susceptibility factor for SSc has been identified. All these findings support the strong autoimmune component underlying SSc pathogenesis and highlight the fact that the development of SSc seems to be determined by shared common genetic and pathogenic mechanisms with other autoimmune diseases and involves specific disease pathways that should be further characterized.

URLS. Illumina iControlDB database, <http://www.illumina.com/science/icontribdb.nlmn/>.

METHODS

Methods and any associated references are available in the online version of the paper at <http://www.nature.com/naturegenetics/>.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

This work was supported by the following grants: T.R.D.J.R. was funded by the VIDI laureate from the Dutch Association of Research (NWO) and Dutch Arthritis Foundation (National Reumafonds). J.M. was funded by GEN-FER from the Spanish Society of Rheumatology, SAF2009-11110 from the Spanish Ministry of Science, CTS-4977 from Junta de Andalucía, Spain and in part by Redes Temáticas de Investigación Cooperativa Sanitaria Program, RD08/0075 (RIER) from Instituto de Salud Carlos III (ISCIII), Spain (J.M.). R.B. is supported by the 13P Consejo Superior de Investigaciones Científicas program funded by the 'Fondo Social Europeo'. B.Z.A. is supported by the Netherlands Organization for Health Research and Development (ZonMW grant 016.096.121). B.K. is supported by the Dutch Diabetes Research Foundation (grant 2008.40.001) and the Dutch Arthritis Foundation (Reumafonds, grant NR 09-1-408). Genotyping of the Dutch control samples was sponsored by US National Institutes of Mental Health funding, R01 MH078075 (R.O.A.). The German controls were from the PopGen biobank (to B.K.). The PopGen project received infrastructure support through the German Research Foundation excellence cluster 'Inflammation at Interfaces'. The US analyses were supported by the US National Institutes of Health and National Institute of Arthritis and Musculoskeletal Diseases (NIH-NIAMS) R01 AR055258, Two-Stage Genome Wide Association Study in Systemic Sclerosis, (M.D.M.) and by the NIH-NIAMS Center of Research Translation (CORT) in SSc (P50AR054144) (F.C.A.), the NIH-NIAMS SSc Family Registry and DNA Repository (N01-AR-0-2251) (M.D.M.), University of Texas Health Science Center-Houston Center for Clinical and Translational Sciences (Houston Clinical and Translational Science Awards Program) (NIH-National Center for Research Resources 3UL1RR024148) (F.C.A.), NIH-NIAMS K08 Award (K08AR054404) (S.K.A.), SSc Foundation New Investigator Award (S.K.A.).

AUTHOR CONTRIBUTIONS

Study Design: T.R.D.J.R., O.G., B.R., J.-E.M., B.P.C.K., F.C.A., J.M., M.D.M.

Collection of data: T.R.D.J.R., M.J.C., M.C.V., A.E.V., A.J.S., J.C.B., B.A.L., A.-M.H.-V., R.A.O., G.R., N.H., C.P.S., N.O.-C., M.A.G.-G., M.F.G.-E., P.A., J.v.L., A.H., J.W., R.H., V.S., F.d.K., F.H., M.M.C., R.M., P.S., R.W., A.K., H.K., E.d.B., T.W., L.P., L.K., L.B., R.S., J.V., M.H., P.G., J.L.N., F.M.W., L.H., P.C., S.A.

Interpretation and analysis of results: T.R.D.J.R., O.G., B.R., J.-E.M., B.Z.A., R.P.-M., J.Y., Y.H., S.-F.W., R.v.'t.S., P.G., A.T.L., C.I.A., S.K.A., B.P.C.K., J.M., M.D.M., A.I., P.C., S.A., P.K.G.

Critical reading of manuscript: T.R.D.J.R., O.G., B.R., J.-E.M., B.Z.A., J.Y., M.J.C., M.C.V., A.E.V., A.J.S., J.C.B., P.L.C.M.v.R., R.v.S., B.A.L., A.-M.H.-V., G.R., N.H.,

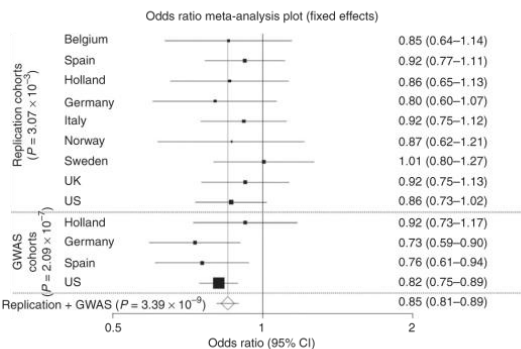


Figure 2 Forest plot showing the odds ratios and confidence intervals of the *CD247* association in the various populations studied in the discovery and replication cohorts.

C.P.S., N.O.-C., M.A.G.-G., M.F.G.-E., P.A., J.V.L., A.H., J.W., R.H., V.S., F.d.K., F.H., M.M.C., R.M., P.S., R.W., A.K., H.K., E.d.B., T.W., L.P., L.B., R.S., J.V., M.H., P.G., C.I.A., J.L.N., F.M.W., L.H., S.K.A., P.G., F.K.T., B.P.C.K., F.C.A., J.M., M.D.M., P.K.G.
Project conception: T.R.D.J.R., B.P.C.K., F.C.A., J.M., M.D.M.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at <http://www.nature.com/naturegenetics/>.

Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions/>.

- Gabrielli, A., Avedimento, E.V. & Krieg, T. Scleroderma. *N. Engl. J. Med.* **360**, 1989–2003 (2009).
- Jimenez, S.A. & Derk, C.T. Following the molecular pathways toward an understanding of the pathogenesis of systemic sclerosis. *Ann. Intern. Med.* **140**, 37–50 (2004).
- Agarwal, S.K., Tan, F.K. & Arnett, F.C. Genetics and genomic studies in scleroderma (systemic sclerosis). *Rheum. Dis. Clin. North Am.* **34**, 17–40 (2008).
- Arnett, F.C. *et al.* Major Histocompatibility Complex (MHC) class II alleles, haplotypes, and epitopes which confer susceptibility or protection in the fibrosing autoimmune disease systemic sclerosis: analyses in 1300 Caucasian, African-American and Hispanic cases and 1000 controls. *Ann. Rheum. Dis.* published online (12 July 2009).
- Rueda, B. *et al.* The *STAT4* gene influences the genetic predisposition to systemic sclerosis phenotype. *Hum. Mol. Genet.* **18**, 2071–2077 (2009).
- Tsuchiya, N. *et al.* Association of *STAT4* polymorphism with systemic sclerosis in a Japanese population. *Ann. Rheum. Dis.* **68**, 1375–1376 (2009).
- Ito, I. *et al.* Association of a functional polymorphism in the *IRF5* region with systemic sclerosis in a Japanese population. *Arthritis Rheum.* **60**, 1845–1850 (2009).
- Dieudé, P. *et al.* Association between the *IRF5* rs2004640 functional polymorphism and systemic sclerosis: a new perspective for pulmonary fibrosis. *Arthritis Rheum.* **60**, 225–233 (2009).
- Gregersen, P.K. & Olsson, L.M. Recent advances in the genetics of autoimmune disease. *Annu. Rev. Immunol.* **27**, 363–391 (2009).
- Hunter, D.J. *et al.* A genome-wide association study identifies alleles in *FGFR2* associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.* **39**, 870–874 (2007).
- Yeager, M. *et al.* Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.* **39**, 645–649 (2007).
- Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455**, 232–236 (2008).
- Krawczak, M. *et al.* PopGen: population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. *Community Genet.* **9**, 55–61 (2006).
- Wellcome Trust Case-Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
- Call, M.E. & Wucherpfennig, K.W. Molecular mechanisms for the assembly of the T cell receptor-CD3 complex. *Mol. Immunol.* **40**, 1295–1305 (2004).
- Krishnan, S. *et al.* Increased caspase-3 expression and activity contribute to reduced CD3zeta expression in systemic lupus erythematosus T cells. *J. Immunol.* **175**, 3417–3423 (2005).
- Krishnan, S. *et al.* Generation and biochemical analysis of human effector CD4 T cells: alterations in tyrosine phosphorylation and loss of CD3 ζ expression. *Blood* **97**, 3851–3859 (2001).
- Krishnan, S., Warke, V.G., Nambiar, M.P., Tsokos, G.C. & Farber, D.L. The FcR gamma subunit and Syk kinase replace the CD3 zeta-chain and ZAP-70 kinase in the TCR signaling complex of human effector CD4 T cells. *J. Immunol.* **170**, 4189–4195 (2003).
- Gorman, C.L. *et al.* Polymorphisms in the *CD3Z* gene influence TCRzeta expression in systemic lupus erythematosus patients and healthy controls. *J. Immunol.* **180**, 1060–1070 (2008).
- Warchol, T. *et al.* The *CD3Z* 844 T>A polymorphism within the 3'-UTR of *CD3Z* confers increased risk of incidence of systemic lupus erythematosus. *Tissue Antigens* **74**, 68–72 (2009).

ONLINE METHODS

Subjects. Because SSc is a relatively rare autoimmune disorder (estimated prevalence in populations of European descent ~0.01%), large sets of subjects with SSc can best be recruited through international collaboration. Consequently, to achieve the total of 2,296 SSc cases and 5,171 healthy control individuals analyzed in the present study, we included four case series having participants with European ancestry, from the United States, Spain, Germany and The Netherlands. The cases from the United States (initial $n = 1,678$; after application of quality-control criteria, $n = 1,486$; 179 men, 1,307 women; mean age = 54.5 years (median = 55.0 years); standard deviation (s.d.) = 12.9) were obtained from May 2001 to December 2008 from three US sources—the University of Texas Health Science Center–Houston, The Johns Hopkins University Medical Center and the Fred Hutchinson Cancer Center—with each source enrolling patients from a US-wide catchment area. Whole-genome genotyping data from US control individuals (initial $n = 5,520$) were obtained from the following three publicly available databases: (i) breast cancer controls from the CGEMS studies, (ii) prostate cancer controls from CGEMS and (iii) controls from Illumina iControlDB. After sex-matching and application of quality-control criteria, 419 men and 3,058 women controls were analyzed.

The initial European SSc cases series came from previously established collections with nationally representative recruitment of 380 Spanish, 288 German and 190 Dutch SSc cases. Main demographical and clinical data of European SSc study participants have been described previously^{5,21}. As a control population, healthy unrelated individuals of Spanish (initial $n = 414$), German (initial $n = 678$) and Dutch (initial $n = 643$) origin were included in the study. Whole-genome genotyping data from German controls were from the PopGen Biobank and, for the Dutch controls, were from a previous study^{12,13}.

To further confirm associations found during the GWAS stage, we collected a large independent replication cohort of individuals with European ancestry from Belgium, Spain, Holland, Germany, Italy, Norway, Sweden, the United Kingdom and the United States. A total of 2,753 SSc cases and 4,569 healthy controls were recruited for this second stage (Supplementary Table 4).

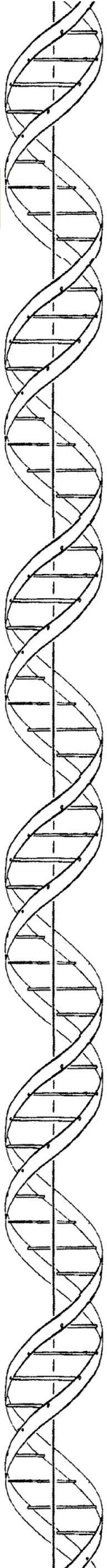
All cases either met the American College of Rheumatology Preliminary criteria for the classification of SSc or had at least three of the five CREST (calcinosis, Raynaud's phenomenon, esophageal dysmotility, sclerodactyly, telangiectasias) features²². Main clinical features of SSc cases are included in Supplementary Table 5.

Collection of blood samples and clinical information from case and control subjects was undertaken with informed consent and relevant ethical review board approval from each contributing center in accordance with the tenets of the Declaration of Helsinki.

Genotyping. The GWAS genotyping of the Spanish SSc cases and controls together with Dutch and German SSc cases was performed at the Department of Medical Genetics of the University Medical Center Utrecht (The Netherlands) using the commercial release Illumina HumanCNV370K BeadChip, which contains 300,000 standard SNPs with an additional 52,167 markers designed to specifically target nearly 14,000 copy number variant regions of the genome, for a total of over 370,000 markers. This system delivers high genomic coverage of the SNPs from Phase I and II of the HapMap Project (see URLs), capturing 81% of the HapMap variation at $r^2 > 0.8$ in European-descended populations. Genotype data for Dutch and German controls were obtained from the Illumina Human 550K BeadChip available from a previous study^{12,13}. The SSc case group from the United States was genotyped at Boas Center for Genomics and Human Genetics, Feinstein Institute for Medical Research, North Shore Long Island Jewish Health System using the Illumina Human610-Quad BeadChip capturing 89% of the HapMap CEU variation at $r^2 > 0.8$. CGEMS and Illumina iControlDB controls were genotyped on the Illumina Hap550K-BeadChip. For the replication phase, SNPs reaching GWAS significance located in new potential SSc susceptibility loci (rs2056626 for *CD247*, rs10515998 for *CDH7* and rs4959270 for *EXOC2-IRF4*) were genotyped in the replication cohorts using Applied Biosystems' TaqMan SNP genotyping Assays on an ABI Prism 7900HT real-time thermocycler. Markers with call rates of 95% or less were excluded, as were markers whose allele distributions deviated strongly from Hardy-Weinberg equilibrium in controls ($P < 10^{-5}$). Only markers with minor allele frequencies of $\geq 1\%$ in both cases and controls were included in the analyses.

Statistical analysis. Statistical analyses were undertaken using R (v2.6), Stata (v8) and PLINK (v1.06) software (see URLs)²³. All reported P values are two-sided. Using PLINK, we identified and excluded pairs of genetically related subjects or duplicates and excluded the genetic-pair members with lower call rate. To identify individuals who might have non-western European ancestry, we merged our case and control data with the data from the HapMap Project (60 western European (CEU), 60 Nigerian (YRI), 90 Japanese (JPT) and 90 Han Chinese (CHB) samples). We used principal component analysis as implemented in HelixTree (see URLs), plotting the first two principal components for each individual. All individuals who did not cluster with the main CEU cluster (defined as deviating more than 4 standard deviations from the cluster centroids) were excluded from subsequent analyses. The principal components derived on the resulting sample look typical for populations of European origin (Supplementary Fig. 3)²⁴. Additionally, we excluded individuals with low call rate (11 individuals from the US group, 24 from the Spanish, 1 from the German and 1 from the Dutch), relatedness (50 from the US group, 2 from the Spanish, 1 from the German and 1 from the Dutch), non-European ancestry (42 from the US group, 5 from the Spanish, 6 from the German and 4 from the Dutch) and inconsistent gender (83 from the US group, 2 from the Spanish, 2 from the German and 2 from the Dutch). Then we filtered for SNP quality, removing SNPs with a genotyping success call rate $< 98\%$ and those showing $MAF < 1\%$. Deviation of the genotype frequencies in the controls from those expected under Hardy-Weinberg equilibrium was assessed by a χ^2 test or Fisher's exact test when an expected cell count was < 5 . SNPs strongly deviating from Hardy-Weinberg equilibrium ($P < 10^{-5}$) were eliminated from the study. For the combined analysis of the four datasets, the same quality controls per individual and per SNP were applied with the exception of the Hardy-Weinberg equilibrium requirement. The genotyping success call rate on the merged dataset after all these quality filters were applied was 99.83% in the GWAS cohorts. In the replication cohorts, genotyping success call rate was 98.16% after quality filtering. The association between each SNP and the risk of scleroderma in each dataset was assessed by the Cochran-Armitage trend test. Odds ratios and associated 95% CIs were calculated using unconditional logistic regression.

To determine if SNPs that were associated at genome-wide significance belonged to extensive linkage disequilibrium (LD) blocks, we investigated the LD pattern (using an r^2 parameter) on a 1-Mb region surrounding significant SNPs (Supplementary Fig. 4a–e). No strong LD ($r^2 > 0.8$) was observed among the investigated SNPs and other variants on the region, except in the case of rs12537284, which was in LD with rs10488631 ($r^2 = 0.82$) in the *TNPO3-IRF5* region, and both were found to be genome-wide significantly associated with other variants on the region (Table 1). The meta-analysis of the four-study series was conducted using standard methods based on the Cochran-Mantel-Haenszel test. A Breslow-Day test was performed for all SNPs to assess the heterogeneity of the effect in different populations. We tested for the population structure and possibility of differential genotyping of cases and controls using quantile-quantile plots of test statistics and we calculated the inflation factor λ by dividing the median of the test statistics s by the expected median from a χ^2 distribution with one degree of freedom. There was evidence of modest inflation of the test statistics ($\lambda = 1.069$ total, or 1.066 after exclusion of the HLA region), indicating a potential effect of the population substructure on the results. We therefore applied a genomic-control correction to our results. Alternatively, we also derived principal components on a population-specific basis using HelixTree software and applied an adjustment for the five first principal components as well as gender separately for each country using logistic regression, after which we combined the effects for each SNP by meta-analysis using inverse variance method (corresponding P values are presented in Table 1 and Supplementary Table 2). The results from this analysis were consistent with the results from the genomic-control-corrected Mantel-Haenszel meta-analysis. We then proceeded to analyze the association of three new SNPs found during the GWAS screen on the replication cohorts. Data were filtered according to same procedures as the GWAS stage. Analysis was carried out by Mantel-Haenszel meta-analysis of all the independent replication cohorts to control for differences between groups. We then did meta-analysis of all the replication and GWAS cohorts for these SNPs using the same Mantel-Haenszel statistical procedure. Results are shown in Table 2.



URLs. HapMap, <http://www.hapmap.org>; R, <http://www.r-project.org/>; PLINK, <http://pngu.mgh.harvard.edu/purcell/plink/>; HelixTree, http://www.goldenhelix.com/SNP_Variation/HelixTree/index.html.

21. Gourh, P. *et al.* Association of the PTPN22 R620W polymorphism with anti-topoisomerase I- and anticentromere antibody-positive systemic sclerosis. *Arthritis Rheum.* **54**, 3945–3953 (2006).

22. Anonymous. Preliminary criteria for the classification of systemic sclerosis (scleroderma). Subcommittee for scleroderma criteria of the American Rheumatism Association Diagnostic and Therapeutic Criteria Committee. *Arthritis Rheum.* **23**, 581–590 (1980).

23. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

24. Tian, C. *et al.* European population genetic substructure: further definition of ancestry informative markers for distinguishing among diverse European ethnic groups. *Mol. Med.* **15**, 371–383 (2009).

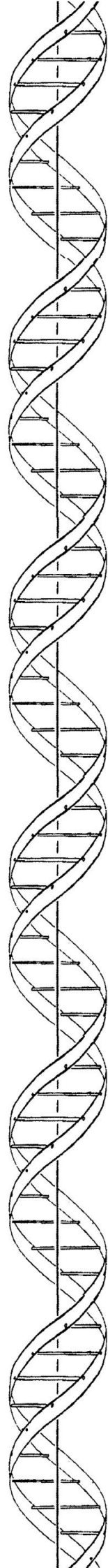
© 2010 Nature America, Inc. All rights reserved.



Supplementary table 1. Study design

Population	GWAS genotyping platform	N Case/Controls After genotyping quality controls	Genotyped SNPs	Analyzed SNPs (overlapping between populations)	Statistical Power*
USA	Illumina Human 550K /(1) breast cancer controls CGEMS, (2) prostate cancer controls CGEMS; and (3) Illumina controls (ref)	1486/3477	488.793	279.621	0.99
Spain	Illumina Human CNV370K BeadChip/ Illumina Human CNV370K BeadChip	364/384	322.967	279.621	0.90
Germany	Illumina Human CNV370K BeadChip/ Illumina Human550K (ref)	270/671	308.299	279.621	0.92
The Netherlands	Illumina Human CNV370K BeadChip/ Illumina Human550K (ref)	176/639	308.349	279.621	0.81
Total	-	2296/5171	-		0.99

*for detecting and OR of 1.5 with a MAF of 0.20



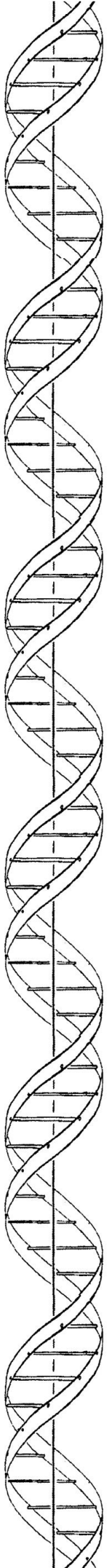
Supplementary table 2. SNPs associated with SSc at genome-wide significance within the MHC region.

CHR	SNP	Closest Gene	BP	Minor Allele	MAF	GC	PC	OR	CI 95%
						Corrected P Value	Corrected P value		
6	rs6457617	HLA-DQB1	32,771,829	C	0.466	3.98E-17	3.85E-19	0.729	0.68-0.78
6	rs2856705	HLA-DQB1	32,778,934	A	0.096	2.19E-14	3.71E-14	0.591	0.52-0.67
6	rs9275390	HLA-DQB1	32,777,134	C	0.271	2.17E-11	1.87E-10	1.314	1.22-1.42
6	rs5000634	HLA-DQB1	32,771,542	C	0.403	3.88E-11	9.13E-13	1.283	1.20-1.38
6	rs443198	NOTCH4	32,298,384	C	0.355	5.98E-10	6.82E-12	0.784	0.73-0.84
6	rs2071286	NOTCH4	32,287,874	A	0.25	2.68E-09	1.67E-10	1.284	1.19-1.39
6	rs2516399	MICB	31,589,278	C	0.105	5.68E-09	6.27E-09	0.688	0.61-0.78
6	rs1521	MICA	31,458,683	C	0.235	7.18E-09	8.96E-09	0.769	0.71-0.84
6	rs9275312	HLA-DQB1	32,773,706	G	0.141	8.04E-09	3.70E-08	1.345	1.22-1.48
6	rs479536	NOTCH4	32,301,656	T	0.058	1.59E-08	1.50E-08	0.615	0.52-0.72
6	rs2071295	TNXB	32,146,678	A	0.34	2.22E-08	2.71E-09	1.242	1.15-1.34
6	rs2523477	MICA	31,468,368	G	0.082	2.64E-08	1.94E-07	0.669	0.58-0.77
6	rs12153855	TNXB	32,182,782	C	0.098	2.70E-08	5.08E-08	0.692	0.61-0.78
6	rs2844494	MICB	31,591,394	G	0.312	2.99E-08	3.22E-08	0.798	0.74-0.86
6	rs2239689	TNXB	32,138,262	T	0.342	4.01E-08	4.64E-09	1.237	1.15-1.33
6	rs3129871	HLA-DRA	32,514,320	A	0.343	4.92E-08	4.95E-08	0.806	0.75-0.87
6	rs1035798	AGER	32,259,200	T	0.278	5.32E-08	2.66E-09	1.249	1.16-1.35
6	rs3104398	HLA-DQA2	32,793,663	A	0.11	8.16E-08	2.81E-07	0.718	0.64-0.81
6	rs2261033	BAT2	31,711,570	C	0.49	9.83E-08	2.49E-06	1.219	1.14-1.31
6	rs204999	PRRT1	32,217,957	G	0.26	1.13E-07	2.30E-09	0.795	0.73-0.86
6	rs9277554	HLA-DPB1	33,163,516	T	0.297	1.33E-07	4.64E-08	1.236	1.15-1.33
6	rs2516398	MICB	31,589,505	C	0.312	1.53E-07	1.64E-07	0.807	0.75-0.87
6	rs6901221	HLA-DPB2	33,206,254	C	0.166	1.91E-07	1.19E-07	1.29	1.18-1.41
6	rs7774954	HLA-DQB2	32,832,167	A	0.066	2.13E-07	3.06E-07	0.662	0.57-0.77
6	rs2248462	MICB	31,554,775	A	0.207	2.22E-07	3.27E-07	0.784	0.72-0.86
6	rs2596480	MICA	31,533,964	A	0.077	2.40E-07	6.55E-07	0.683	0.59-0.79
6	rs707939	MSH5	31,834,667	T	0.37	2.43E-07	6.55E-07	1.217	1.13-1.31
6	rs2516509	MICB	31,557,973	G	0.207	3.08E-07	4.19E-07	0.786	0.72-0.86
6	rs2075800	HSPA1L	31,885,925	A	0.355	3.70E-07	1.58E-07	1.216	1.13-1.31
6	rs3095352	DDR1	30,913,900	G	0.424	4.19E-07	3.12E-06	0.826	0.77-0.89
6	rs12665700	C6orf205	31,104,111	T	0.137	8.18E-07	7.25E-07	1.293	1.17-1.43
6	rs3129941	C6orf10	32,445,664	A	0.227	9.50E-07	1.43E-06	0.801	0.74-0.87
6	rs6941112	STK19	32,054,593	A	0.343	1.24E-06	2.29E-07	1.206	1.12-1.30

Supplementary Table 3. Association analysis of overall top SNPs on each population separately. Right most column show Breslow-Day P value for the heterogeneity on the odds ratio.

CHR	Gene	SNP	Position	Minor Allele	Population	MAF (case/control)	P Value*	OR (95 CI)	Breslow-day P value
6p21	HLA-DQB1	rs6457617	32771829	C	United States	0.403/0.487	1.17E-14	0.71 (0.65-0.77)	5.11E-02
				C	Spain	0.402/0.474	5.22E-03	0.74 (0.61-0.92)	
				C	Germany	0.400/0.504	4.67E-05	0.66 (0.54-0.80)	
				C	Holland	0.491/0.488	9.14E-01	1.01 (0.80-1.28)	
7q32	TNPO3/ IRF5	rs10488631	128381419	C	United States	0.149/0.105	5.44E-10	1.49 (1.13-1.69)	4.34E-01
				C	Spain	0.115/0.090	1.03E-01	1.32 (0.94-1.85)	
				C	Germany	0.170/0.101	2.62E-05	1.84 (1.38-2.44)	
				C	Holland	0.125/0.095	9.69E-02	1.36 (0.94-1.97)	
		rs12537284	128505142	A	United States	0.169/0.132	1.27E-06	1.34 (1.19-1.51)	5.80E-01
				A	Spain	0.143/0.132	5.24E-01	1.10 (0.82-1.48)	
				A	Germany	0.161/0.125	3.94E-02	1.34 (1.01-1.78)	
				A	Holland	0.145/0.117	1.52E-01	1.28 (0.91-1.80)	
2q32	STAT4	rs3821236	191611003	A	United States	0.234/0.204	7.74E-04	1.19 (1.08-1.32)	3.97E-02
				A	Spain	0.278/0.189	4.88E-05	1.65 (1.29-2.10)	
				A	Germany	0.257/0.194	2.54E-03	1.44 (1.13-1.82)	
				A	Holland	0.278/0.203	2.39E-03	1.52 (1.16-1.99)	
1q22-23	CD247	rs2056626	165687049	G	United States	0.379/0.428	5.82E-06	0.82 (0.75-0.89)	4.26E-01
				G	Spain	0.324/0.384	1.60E-02	0.77 (0.62-0.95)	
				G	Germany	0.344/0.402	1.95E-02	0.78 (0.63-0.96)	
				G	Holland	0.420/0.425	8.90E-01	0.98 (0.77-1.25)	
6p25	EXOC2/ IRF4	rs4959270	402748	A	United States	0.466/0.502	1.33E-03	0.87 (0.80-0.95)	5.88E-02
				A	Spain	0.408/0.513	4.63E-05	0.65 (0.53-0.80)	
				A	Germany	0.392/0.466	3.54E-03	0.74 (0.60-0.91)	
				A	Holland	0.426/0.470	1.39E-01	0.84 (0.66-1.06)	
18q22	CDH7	rs10515998	61521202	G	United States	0.056/0.039	1.32E-04	1.47 (1.20-1.79)	5.30E-03
				G	Spain	0.060/0.059	9.65E-01	1.01 (0.66-1.55)	
				G	Germany	0.093/0.052	1.20E-03	1.86 (1.27-2.72)	
				G	Holland	0.062/0.018	1.28E-05	3.54 (1.94-6.48)	

* P values uncorrected for λ

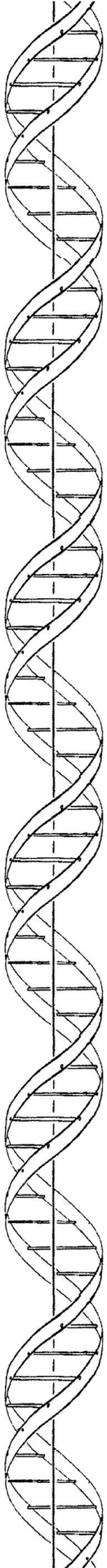


Supplementary table 4. Replication phase study design.

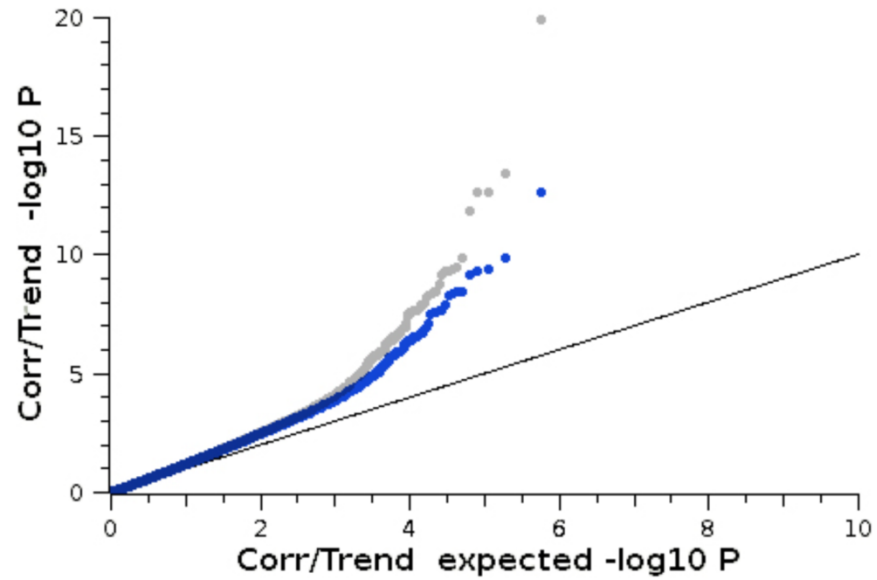
Population	N Case/Controls
Belgium	189/274
Spain	455/739
The Netherlands	228/279
Germany	207/285
Italy	348/727
Norway	113/283
Sweden	279/455
UK	500/384
US	434/1143
Total	2753/4569

Supplementary table 5. Main clinical features of SSc patients included in the study.

	Sex (cases/controls)		Subtype		Anti-centromere		Anti-topoisomerase	
	Female	Male	Difuse	Limited	Positive	Negative	Positive	Negative
Overall	0.80/0.79	0.11/0.21	0.31	0.58	0.32	0.58	0.20	0.71
US	0.88/0.88	0.12/0.12	0.34	0.61	0.29	0.62	0.16	0.77
Spain	0.88/0.75	0.10/0.25	0.25	0.59	0.42	0.41	0.20	0.61
German	0.88/0.62	0.11/0.38	0.35	0.44	0.42	0.53	0.31	0.62
Dutch	0.72/0.51	0.28/0.49	0.13	0.51	0.22	0.69	0.26	0.65
Replication SSc cohorts								
Belgium	0.73/0.45	0.21/0.54	0.31	0.63	0.23	0.47	0.18	0.52
Spain	0.87/0.60	0.11/0.37	0.32	0.59	0.29	0.61	0.14	0.75
The Netherlands	0.71/0.45	0.20/0.54	0.17	0.51	0.22	0.62	0.28	0.64
Germany	0.81/0.44	0.13/0.27	0.37	0.45	0.27	0.35	0.21	0.41
Italy	0.76/0.65	0.07/0.35	0.20	0.60	0.31	0.53	0.32	0.52
Norway	0.85/0.33	0.15/0.66	0.35	0.65	0.53	0.46	0.14	0.84
Sweden	0.78/0.78	0.22/0.22	0.28	0.72	0.27	0.73	0.17	0.83
UK	0.83/0.44	0.16/0.56	0.27	0.72	0.28	0.54	0.11	0.70
US	0.88/0.44	0.12/0.56	0.37	0.50	0.32	0.66	0.12	0.85

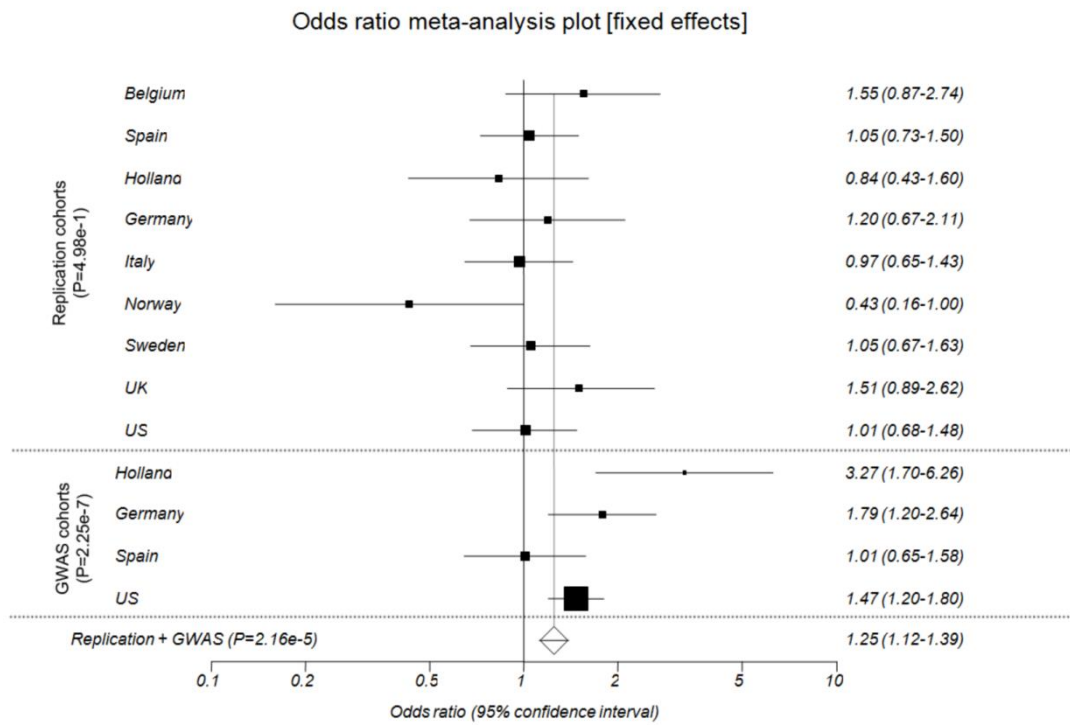


Supplementary figure 1. Quantile-quantile (QQ) plot of the observed P values for association. The grey dots represent the totality of SNPs analyzed ($\lambda = 1.069$) and the blue dots represent all the SNPs analyzed excluding the MHC region ($\lambda = 1.066$).

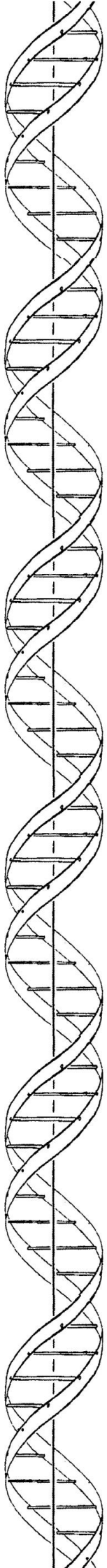
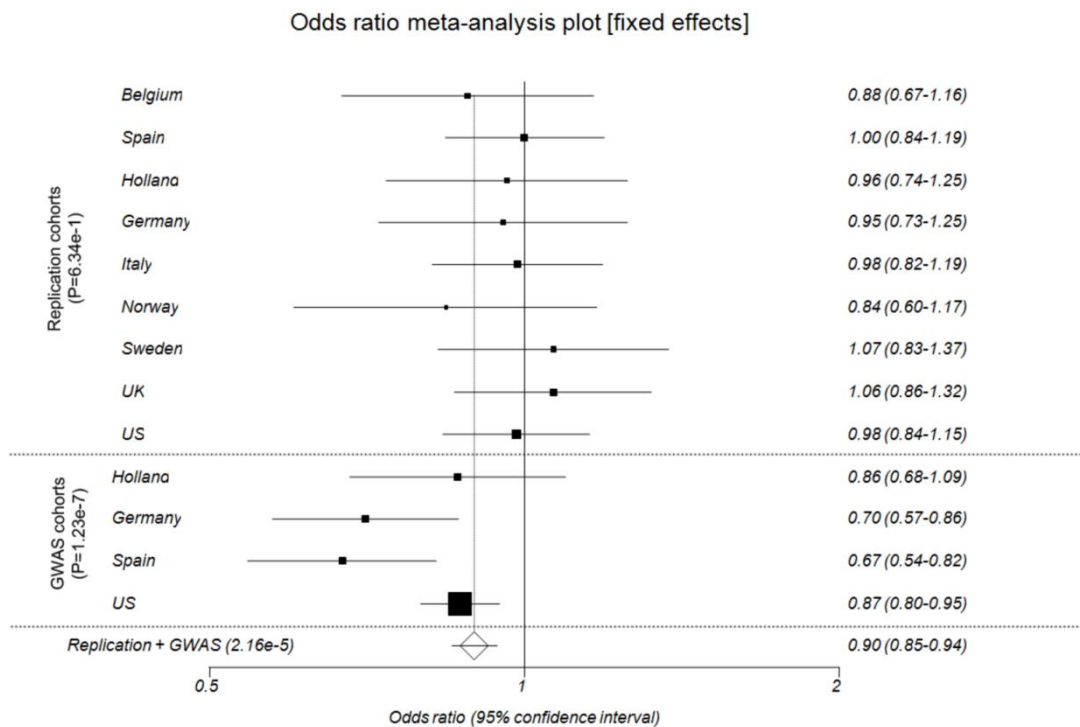


Supplementary figure 2. Forest plot of the *CDH7* and *EXOC2/IRF4* loci at different study stages.

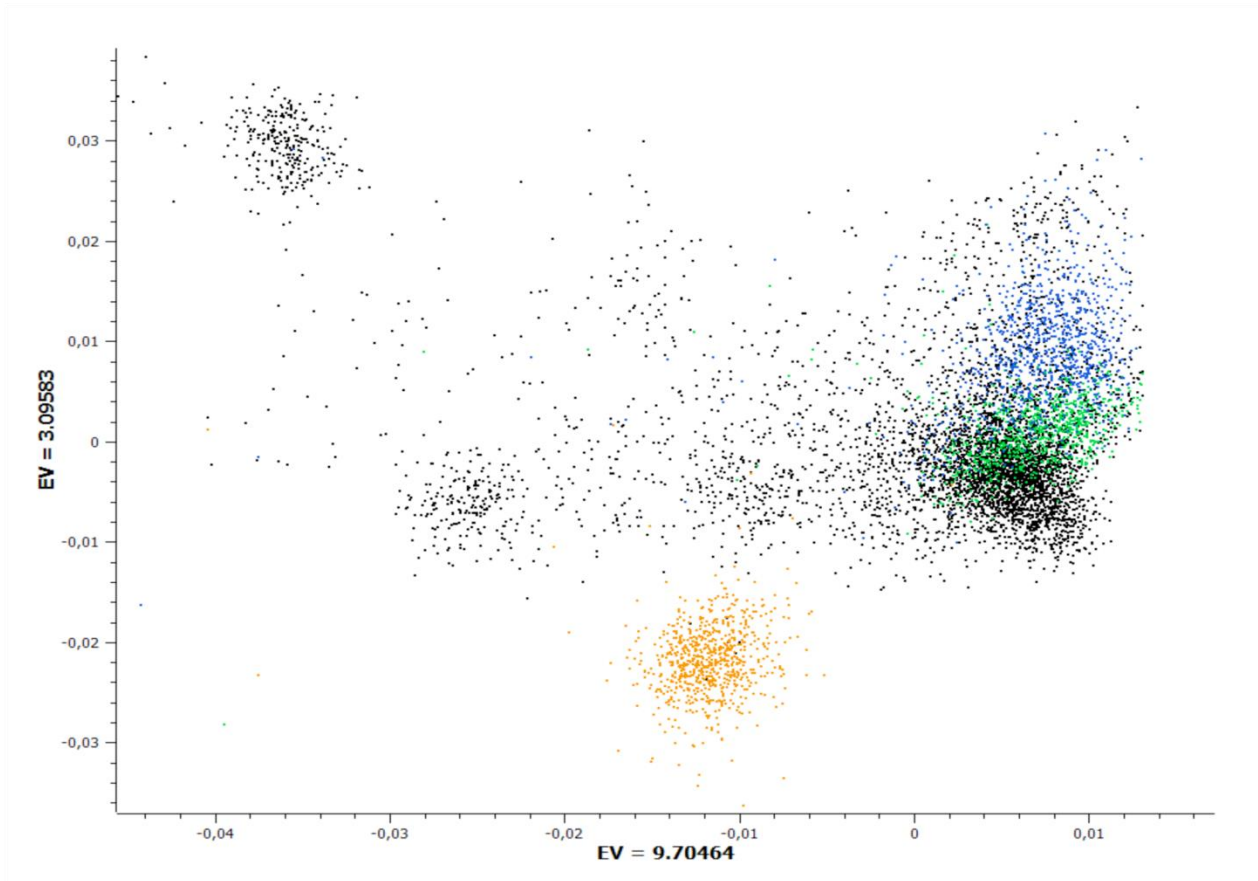
A) *CDH7*



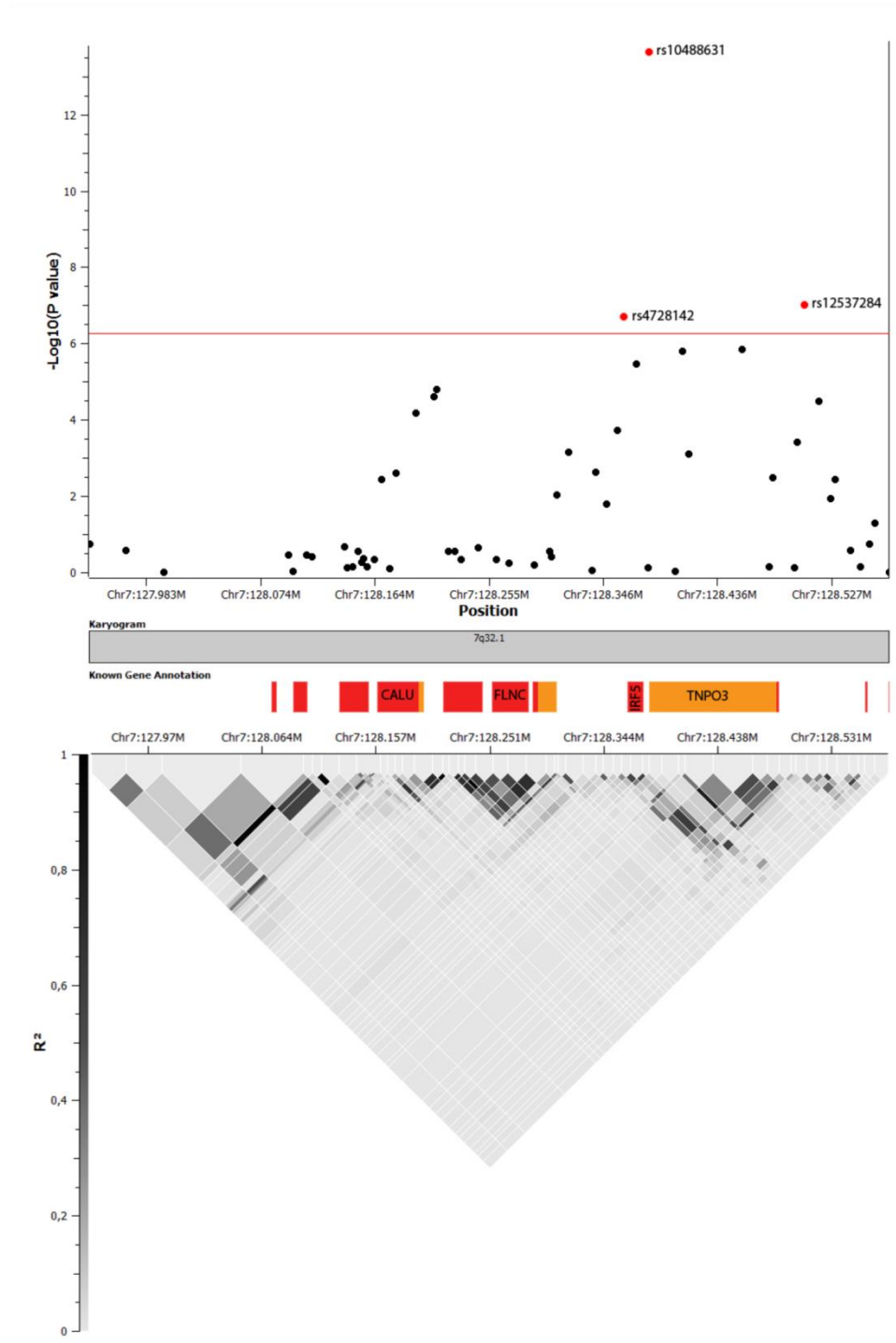
B) *EXOC2/IRF4*



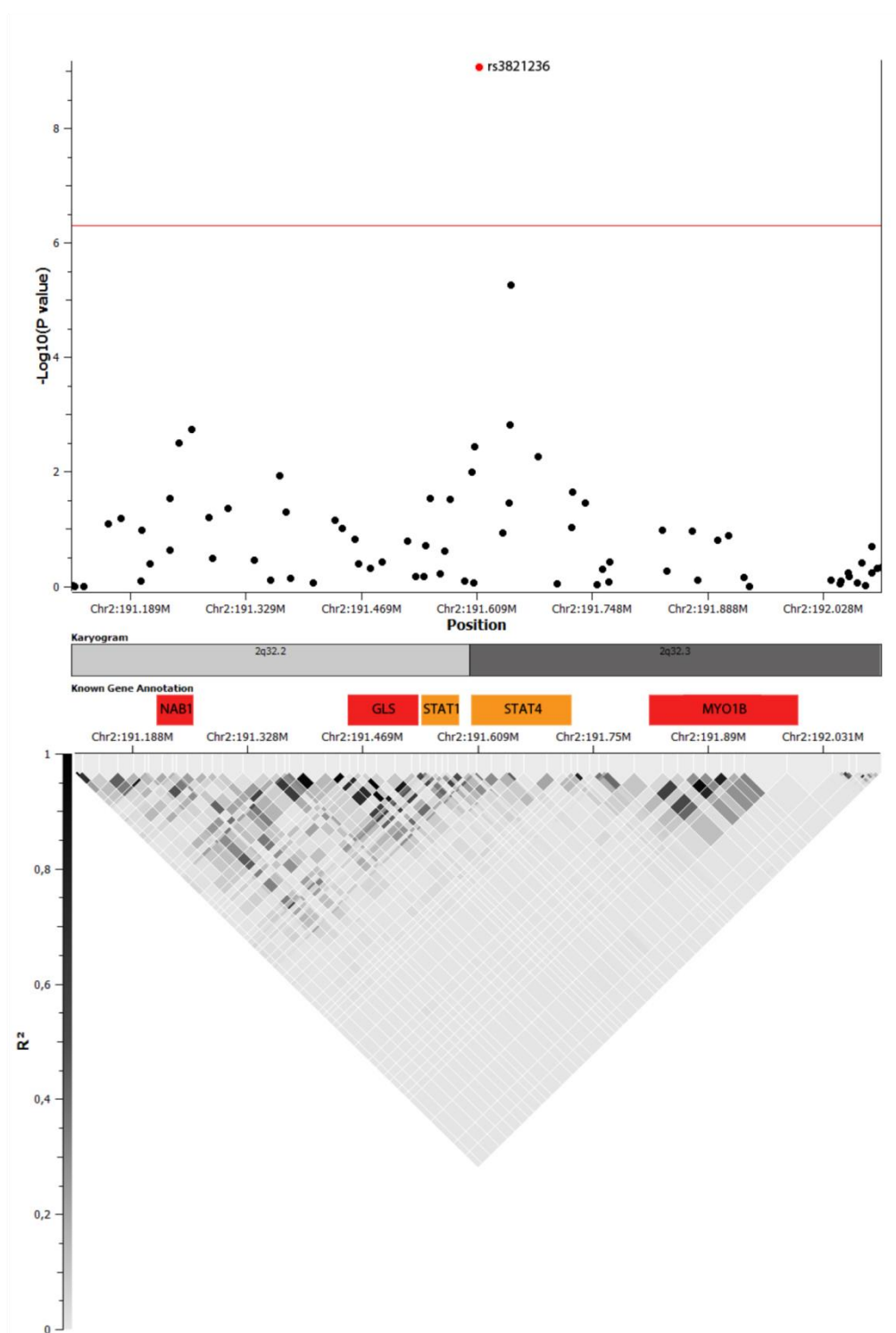
Supplementary figure 3. Principal component plot for the first 2 eigenvectors. US population is represented on black, German on blue, Dutch on green and Spanish on orange.



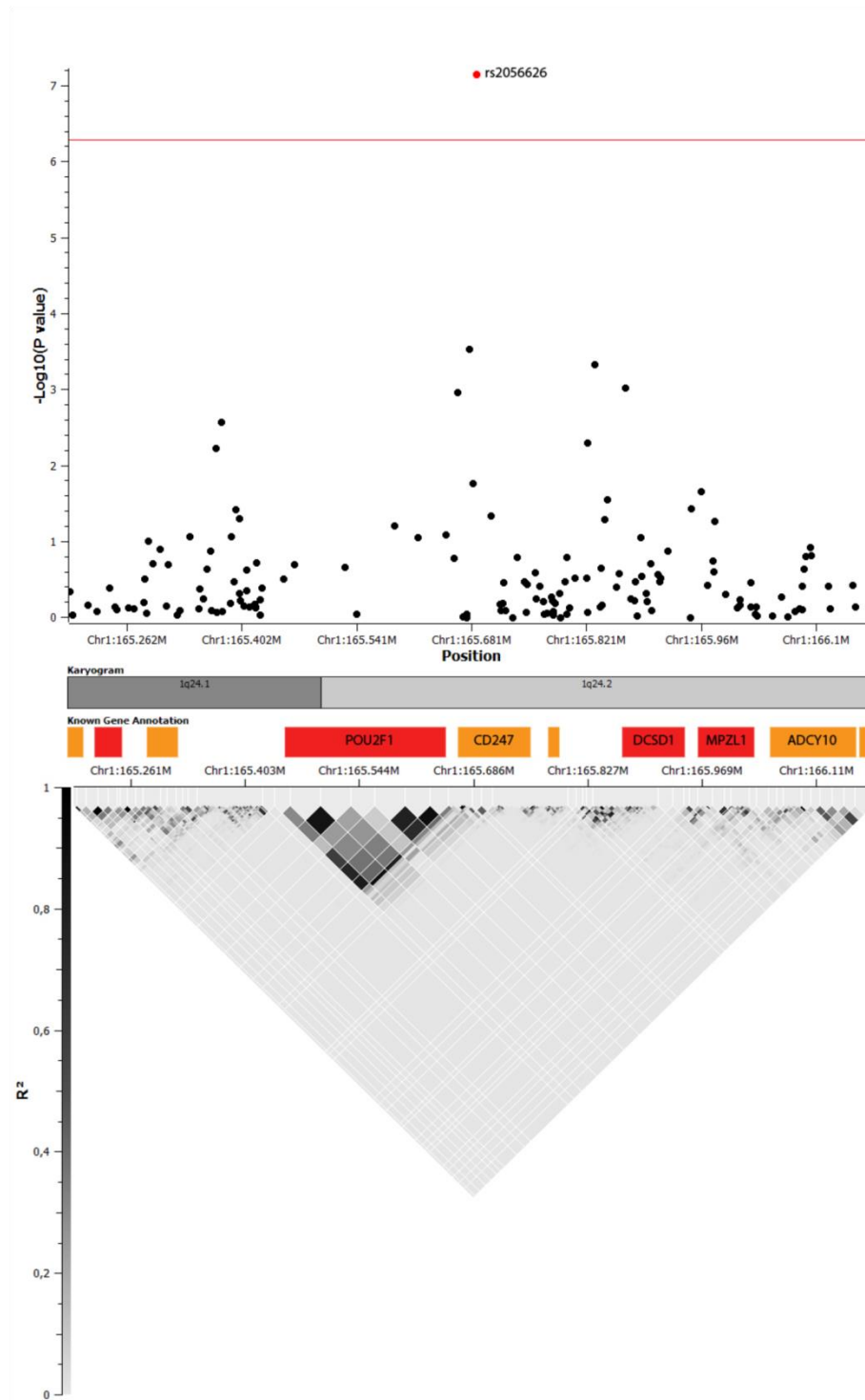
Supplementary figure 4a. Association and linkage disequilibrium (LD) plot on TNPO3/IRF5 association region. P values are uncorrected for λ . LD represented is R^2 . Genome wide significance level is marked with a red line.



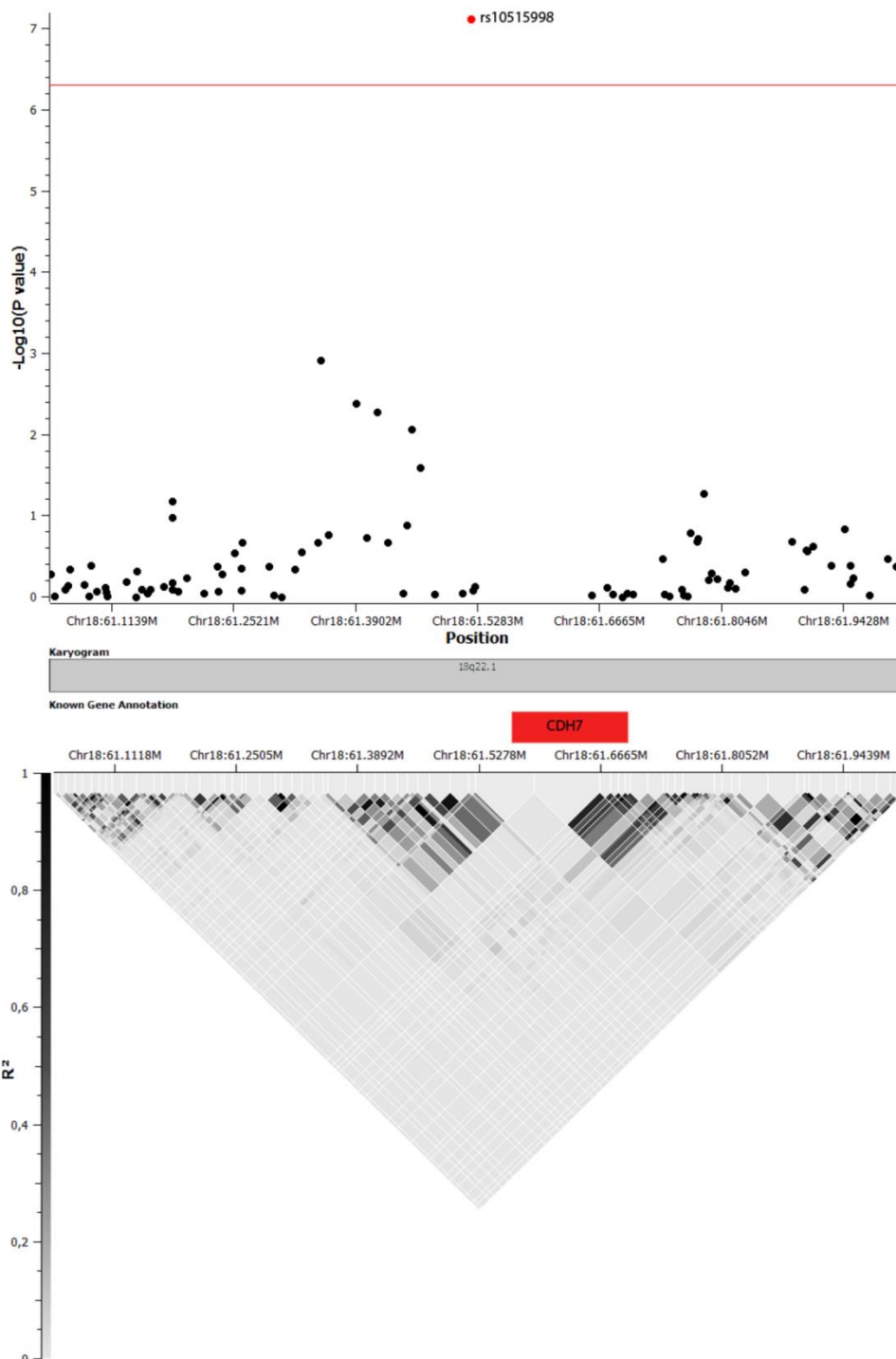
Supplementary figure 4b. Association and linkage disequilibrium (LD) plot on STAT4 association region. P values are uncorrected for λ . LD represented is R^2 . Genome wide significance level is marked with a red line.



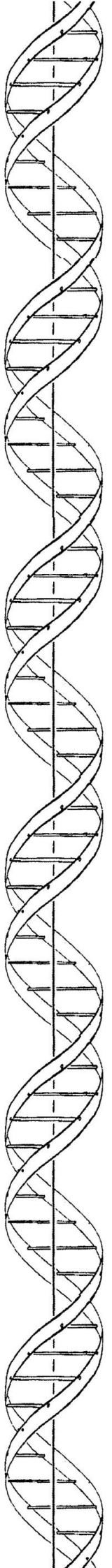
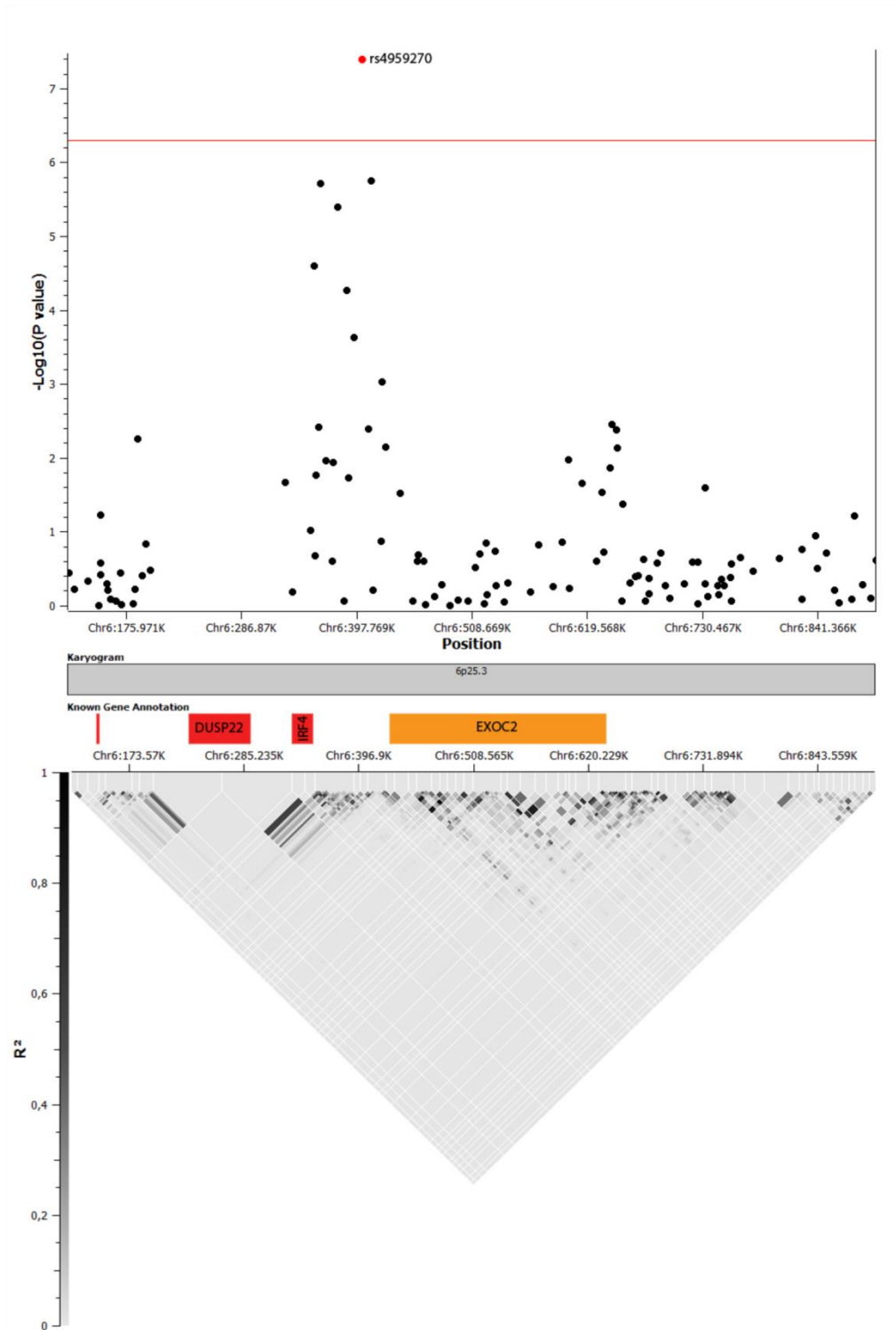
Supplementary figure 4c. Association and linkage disequilibrium (LD) plot on CD247 association region. P values are uncorrected for λ . LD represented is R^2 . Genome wide significance level is marked with a red line.

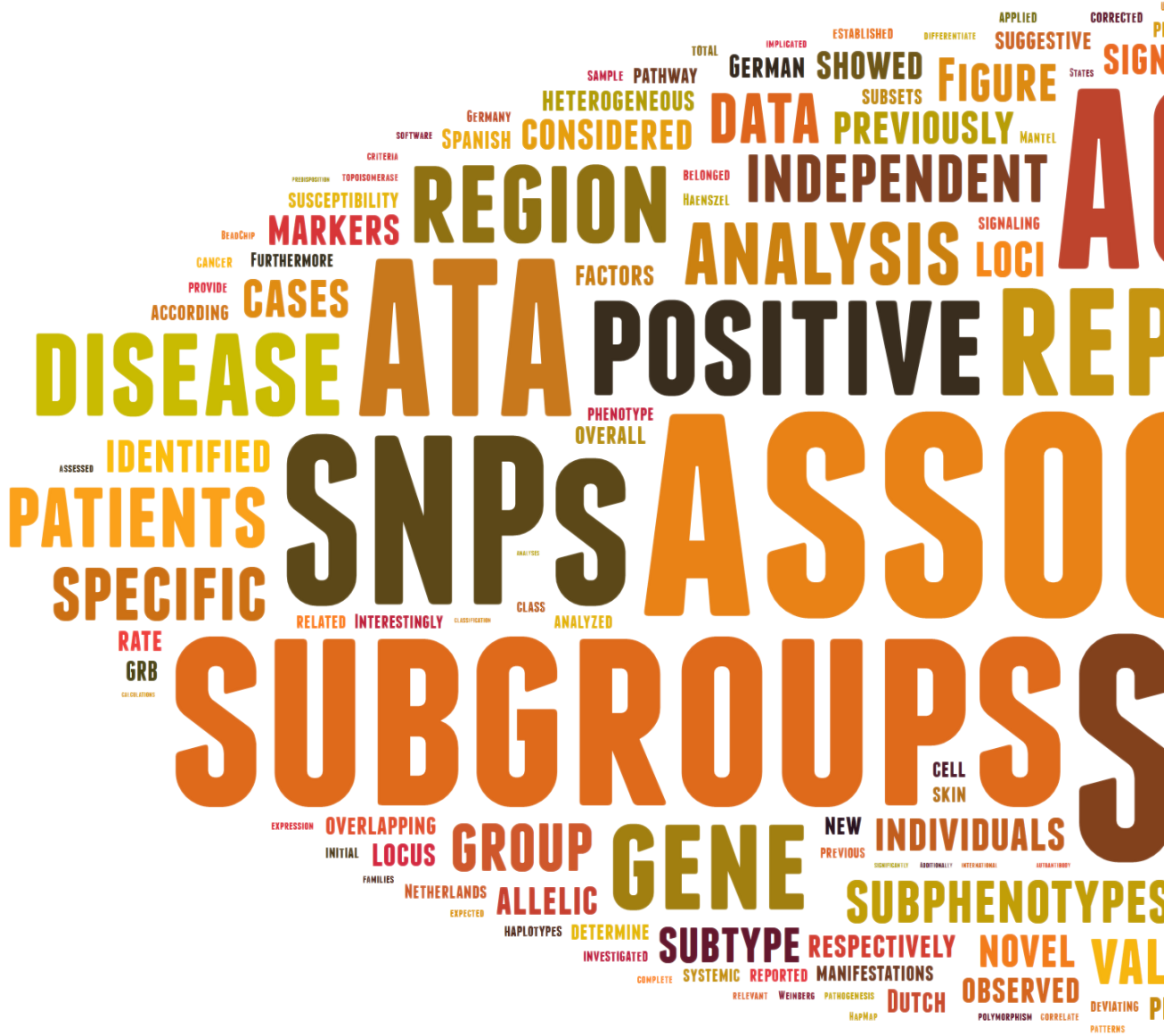
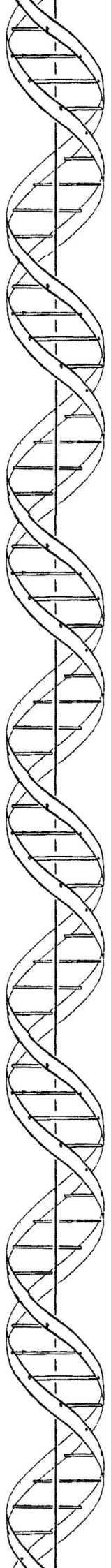


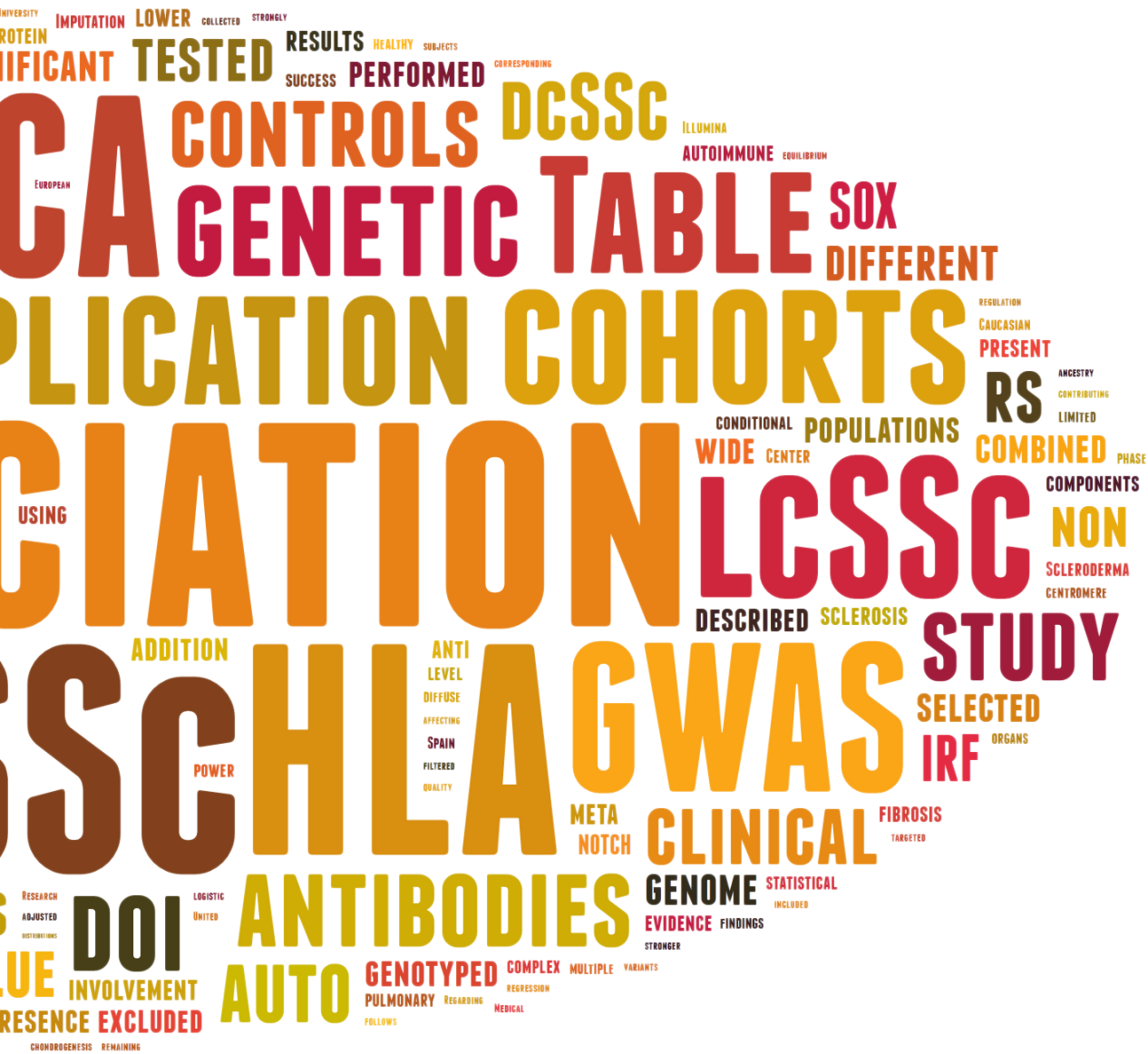
Supplementary figure 4d. Association and linkage disequilibrium (LD) plot on CDH7 association region. P values are uncorrected for λ . LD represented is R^2 . Genome wide significance level is marked with a red line.



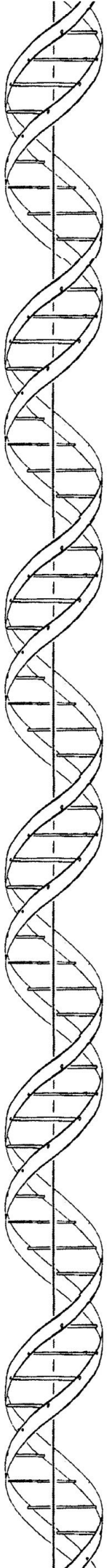
Supplementary figure 4e. Association and linkage disequilibrium (LD) plot on EXOC2/IRF4 association region. P values are uncorrected for λ . LD represented is R^2 . Genome wide significance level is marked with a red line.

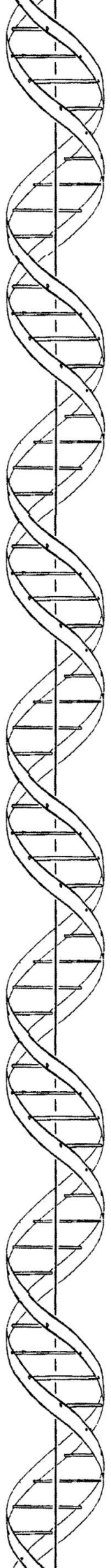






IDENTIFICATION OF NOVEL GENETIC MARKERS ASSOCIATED WITH CLINICAL PHENOTYPES OF SYSTEMIC SCLEROSIS THROUGH A GENOME-WIDE ASSOCIATION STRATEGY. *PLoS GENETICS*, 2011.





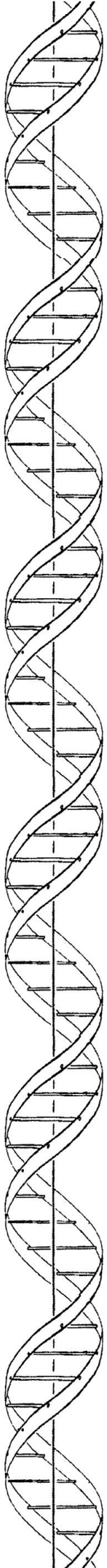
Identification of Novel Genetic Markers Associated with Clinical Phenotypes of Systemic Sclerosis through a Genome-Wide Association Strategy

Olga Gorlova^{1,9*}, Jose-Ezequiel Martin^{2,9}, Blanca Rueda^{2,9}, Bobby P. C. Koeleman^{3,9}, Jun Ying¹, Maria Teruel², Lina-Marcela Diaz-Gallo², Jasper C. Broen⁴, Madelon C. Vonk⁴, Carmen P. Simeon⁵, Behrooz Z. Alizadeh⁶, Marieke J. H. Coenen⁷, Alexandre E. Voskuyl⁸, Annemie J. Schuerwegh⁹, Piet L. C. M. van Riel⁴, Marie Vanthuyne¹⁰, Ruben van 't Slot³, Annet Italiaander³, Roel A. Ophoff³, Nicolas Hunzelmann¹¹, Vicente Fonollosa⁵, Norberto Ortego-Centeno¹², Miguel A. González-Gay¹³, Francisco J. García-Hernández¹⁴, María F. González-Escribano¹⁵, Paolo Airo¹⁶, Jacob van Laar¹⁷, Jane Worthington¹⁸, Roger Hesselstrand¹⁹, Vanessa Smith²⁰, Filip de Keyser²⁰, Fredric Houssiau¹⁰, Meng May Chee²¹, Rajan Madhok²¹, Paul G. Shiels²², Rene Westhovens²³, Alexander Kreuter²⁴, Elfride de Baere²⁵, Torsten Witte²⁶, Leonid Padyukov²⁷, Annika Nordin²⁷, Raffaella Scorza²⁸, Claudio Lunardi²⁹, Benedicte A. Lie³⁰, Anna-Maria Hoffmann-Vold³¹, Øyvind Palm³¹, Paloma García de la Peña³², Patricia Carreira³³, Spanish Scleroderma Group^a, John Varga³⁴, Monique Hinchcliff³⁴, Annette T. Lee³⁵, Pravitt Gourh³⁶, Christopher I. Amos¹, Frederick M. Wigley³⁷, Laura K. Hummers³⁸, J. Hummers³⁷, J. Lee Nelson³⁸, Gabriella Riemekasten³⁹, Ariane Herrick¹⁸, Lorenzo Beretta²⁸, Carmen Fonseca⁴⁰, Christopher P. Denton⁴⁰, Peter K. Gregersen³⁵, Sandeep Agarwal³⁶, Shervin Assassi³⁶, Filemon K. Tan³⁶, Frank C. Arnett^{36†}, Timothy R. D. J. Radstake^{4†}, Maureen D. Mayes^{36†}, Javier Martin^{2†*}

1 Department of Epidemiology, M. D. Anderson Cancer Center, Houston, Texas, United States of America, **2** Instituto de Parasitología y Biomedicina López-Neyra, Consejo Superior de Investigaciones Científicas, Granada, Spain, **3** Department of Medical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands, **4** Department of Rheumatology, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands, **5** Servicio de Medicina Interna, Hospital Valle de Hebrón, Barcelona, Spain, **6** University Medical Centre Groningen, Department of Epidemiology, Groningen, The Netherlands, **7** Department of Human Genetics, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands, **8** VU University Medical Center, Amsterdam, The Netherlands, **9** Department of Rheumatology, University of Leiden, Leiden, The Netherlands, **10** Cliniques Universitaires Saint-Luc, Université Catholique de Louvain, Brussels, Belgium, **11** Department of Dermatology, University of Cologne, Cologne, Germany, **12** Servicio de Medicina Interna, Hospital Clínico Universitario, Granada, Spain, **13** Servicio de Reumatología, Hospital Marqués de Valdecilla, Santander, Spain, **14** Servicio de Medicina Interna, Hospital Virgen del Rocío, Sevilla, Spain, **15** Servicio de Inmunología, Hospital Virgen del Rocío, Sevilla, Spain, **16** Rheumatology Unit and Chair, Spedali Civili, Università degli Studi, Brescia, Italy, **17** Institute of Cellular Medicine, Newcastle University, Newcastle Upon Tyne, United Kingdom, **18** Department of Rheumatology and Epidemiology, University of Manchester, Manchester Academic Health Science Centre, Manchester, United Kingdom, **19** Department of Clinical Sciences, Division of Rheumatology, Lund University, Lund, Sweden, **20** Ghent University, Ghent, Belgium, **21** Centre for Rheumatic Diseases, Glasgow Royal Infirmary Glasgow, United Kingdom, **22** Department of Surgery, Western Infirmary Glasgow, University of Glasgow, Glasgow, United Kingdom, **23** Katholieke Universiteit Leuven, Leuven, Belgium, **24** Department of Dermatology, Josefs-Hospital, Ruhr University Bochum, Germany, **25** Center for Medical Genetics, Ghent University Hospital, Ghent, Belgium, **26** Hannover Medical School, Hannover, Germany, **27** Center for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden, **28** Referral Center for Systemic Autoimmune Diseases, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico and University of Milan, Milan, Italy, **29** Department of Medicine, Policlinico GB Rossi, University of Verona, Italy, **30** Institute of Immunology, Oslo University Hospital Rikshospitalet, Oslo, Norway, **31** Department of Rheumatology, Rikshospitalet, Oslo University Hospital, Oslo, Norway, **32** Servicio de Reumatología, Hospital Ramón y Cajal, Madrid, Spain, **33** Hospital 12 de Octubre, Madrid, Spain, **34** Northwestern University Feinberg School of Medicine, Chicago, Illinois, United States of America, **35** Feinstein Institute of Medical Research, Manhasset, New York, United States of America, **36** The University of Texas Health Science Center–Houston, Houston, Texas, United States of America, **37** The Johns Hopkins University Medical Center, Baltimore, Maryland, United States of America, **38** Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America, **39** Department of Rheumatology and Clinical Immunology, Charité University Hospital, Berlin, Germany, **40** Centre for Rheumatology, Royal Free and University College School, London, United Kingdom

Abstract

The aim of this study was to determine, through a genome-wide association study (GWAS), the genetic components contributing to different clinical sub-phenotypes of systemic sclerosis (SSc). We considered limited (lcSSc) and diffuse (dcSSc) cutaneous involvement, and the relationships with presence of the SSc-specific auto-antibodies, anti-centromere (ACA), and anti-topoisomerase I (ATA). Four GWAS cohorts, comprising 2,296 SSc patients and 5,171 healthy controls, were meta-analyzed looking for associations in the selected subgroups. Eighteen polymorphisms were further tested in nine independent cohorts comprising an additional 3,175 SSc patients and 4,971 controls. Conditional analysis for associated SNPs in the HLA region was performed to explore their independent association in antibody subgroups. Overall analysis showed that non-HLA polymorphism rs11642873 in *IRF8* gene to be associated at GWAS level with lcSSc ($P = 2.32 \times 10^{-12}$, OR = 0.75). Also, rs12540874 in *GRB10* gene ($P = 1.27 \times 10^{-6}$, OR = 1.15) and rs11047102 in *SOX5* gene ($P = 1.39 \times 10^{-7}$, OR = 1.36) showed a suggestive association with lcSSc and ACA subgroups respectively. In the HLA region, we observed highly associated allelic combinations in the *HLA-DQB1* locus with ACA ($P = 1.79 \times 10^{-61}$, OR = 2.48), in the *HLA-DPA1/B1* loci with ATA ($P = 4.57 \times 10^{-76}$, OR = 8.84), and in *NOTCH4* with ACA ($P = 8.84 \times 10^{-21}$, OR = 0.55) and ATA ($P = 1.14 \times 10^{-8}$,



OR = 0.54). We have identified three new non-HLA genes (*IRF8*, *GRB10*, and *SOX5*) associated with SSc clinical and auto-antibody subgroups. Within the HLA region, *HLA-DQB1*, *HLA-DPA1/B1*, and *NOTCH4* associations with SSc are likely confined to specific auto-antibodies. These data emphasize the differential genetic components of subphenotypes of SSc.

Citation: Gorlova O, Martin J-E, Rueda B, Koeleman BPC, Ying J, et al. (2011) Identification of Novel Genetic Markers Associated with Clinical Phenotypes of Systemic Sclerosis through a Genome-Wide Association Strategy. *PLoS Genet* 7(7): e1002178. doi:10.1371/journal.pgen.1002178

Editor: Mark I. McCarthy, University of Oxford, United Kingdom

Received December 16, 2010; **Accepted** May 25, 2011; **Published** July 14, 2011

Copyright: © 2011 Gorlova et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by the following grants: J Martin was funded by GEN-FER from the Spanish Society of Rheumatology, SAF2009-11110 from the Spanish Ministry of Science, CTS-4977 from Junta de Andalucía, Spain, and in part by Redes Temáticas de Investigación Cooperativa Sanitaria Program, RD08/0075 (RIER) from Instituto de Salud Carlos III (ISCIII), Spain. TRDJ Radstake was funded by the VIDI laureate from the Dutch Association of Research (NWO) and Dutch Arthritis Foundation (National Reumafonds). J Martin and TRDJ Radstake were sponsored by the Orphan Disease Program grant from the European League Against Rheumatism (EULAR). BPC Koeleman is supported by the Dutch Diabetes Research Foundation (grant 2008.40.001) and the Dutch Arthritis Foundation (Reumafonds, grant NR 09-1-408). BZ Alizadeh is supported by the Netherlands Organization for Health Research and Development (ZonMW grant 016.096.121). Genotyping of the Dutch control samples was sponsored by US National Institutes of Mental Health funding, R01 MH078075 (ROA). The German controls were from the PopGen biobank (to BPC Koeleman). The PopGen project received infrastructure support through the German Research Foundation excellence cluster "Inflammation at Interfaces." The USA studies were supported by NIH/NIAMS Scleroderma Family Registry and DNA Repository (N01-AR-0-2251), NIH/NIAMS-RO1-AR055258, NIH/NIAMS Center of Research Translation in Scleroderma (1P50AR054144), and the Department of Defense Congressionally Directed Medical Research Programs (W81XWH-07-01-0111). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: oygorlov@mdanderson.org (O Gorlova); martin@ipb.csic.es (J Martin)

✉ These authors contributed equally to this work.

† These authors also contributed equally to this work.

‡ For membership of the Spanish Scleroderma Group, please see Text S1.

Introduction

Genetic factors play an essential role in scleroderma or systemic sclerosis (SSc) etiology as in most complex autoimmune diseases [1]. Multiple reports of well powered candidate gene association and replication studies, together with the first genome-wide association study (GWAS) in this disease have led to the establishment of the Major histocompatibility complex (MHC), *STAT4*, *IRF5*, *BLK*, *BANK1*, *TNFSF4* and *CD247* as SSc susceptibility genes [2–15].

SSc is a clinically heterogeneous disease with a wide range of clinical manifestations, ranging from mild skin fibrosis with minimal internal organ disease to severe skin and organ involvement, reflecting the three main pathological events that characterize this disease: endothelial damage, fibrosis, and autoimmune dysregulation [16]. SSc patients are classified into two clinical subgroups based on the extent of skin involvement, limited SSc (lcSSc) and diffuse SSc (dcSSc) that are associated with different clinical complications and prognoses [17]. Another SSc hallmark is the presence of disease specific and usually mutually exclusive auto-antibodies that correlate both with the extent of skin involvement and the various disease manifestations, such as pulmonary fibrosis and renal crisis [18]. The most common are DNA topoisomerase I (ATA), and anti-centromere antibodies (CENP A and/or B proteins) [19]. Each of these auto-antibodies is a marker for relatively distinct clinical subgroups of SSc, with anti-centromere typically associated with limited cutaneous disease, uncommon pulmonary fibrosis, late-onset pulmonary hypertension but generally an overall good prognosis, while ATA is a marker for diffuse skin disease and clinically significant pulmonary fibrosis with a resultant poorer prognosis.

It has been observed that certain SSc clinical features and the presence of disease specific auto-antibodies vary in different countries and ethnicities [20]. This fact supports the likelihood that genetic factors may influence the different clinical features of the

disease and auto-antibody subsets [19]. Furthermore, the affected members within multicase SSc families tend to be concordant for SSc-specific auto-antibodies and HLA haplotypes, thus, providing further evidence for a genetic basis for auto-antibody expression in SSc [21]. Moreover, several studies have reported that certain SSc genetic risk factors correlate with specific clinical subsets of the disease or SSc-related auto-antibodies [4,12,22,23].

In this study, we aimed to identify novel genetic factors associated with different SSc clinical and auto-antibody subsets through a stratified re-analysis of results from a previous GWAS from our group and validation in a large replication study.

Results

First, the genetic associations were tested in each of the four subgroups considered for this study (lcSSc, dcSSc, ACA positive and ATA positive) by the means of χ^2 tests in the GWAS data (individuals from the United States, Spain, Germany and The Netherlands), correcting the *P* values for the genomic inflation factor λ of each subgroup (Figures S1, S2, S3, S4 and Tables S1, S2, S3, S4). We found a total of eighteen novel non-HLA loci associated in these subgroups with a *P* value lower than 1×10^{-5} , seven in the lcSSc subtype, five in the dcSSc subtype, two in ACA positives and four in ATA positives. Next, we proceeded to replicate these associations in nine independent cohorts (from US, Spain, Germany, The Netherlands, Belgium, Italy, Sweden, United Kingdom and Norway). The statistically significant results observed in the replication step are shown in Table 1. The complete set of data is shown in Tables S1, S2, S3 S4.

In addition, exhaustive analysis was performed in the HLA region (megabases 28 to 34 in chromosome 6) with the GWAS data in order to find specific subgroup associations in this region. Due to the fact that most associations found herein in the MHC region have been previously described, we did not perform a replication phase of these findings. Instead, let these results be the

Author Summary

Scleroderma or systemic sclerosis is a complex autoimmune disease affecting one individual of every 100,000 in Caucasian populations. Even though current genetic studies have led to better understanding of the pathogenesis of the disease, much remains unknown. Scleroderma is a heterogeneous disease, which can be subdivided according to different criteria, such as the involvement of organs and the presence of specific autoantibodies. Such subgroups present more homogeneous genetic groups, and some genetic associations with these manifestations have already been described. Through reanalysis of a genome-wide association study data, we identify three novel genes containing genetic variations which predispose to subphenotypes of the disease (*IRF8*, *GRB10*, and *SOX5*). Also, we better characterize the patterns of associated loci found in the HLA region. Together, our findings lead to a better understanding of the genetic component of scleroderma.

replication for previous works. It is also noteworthy that all independent associations found within the MHC region have almost exactly the same ORs in the four GWAS cohorts separately, thus, replicating themselves.

Clinical Manifestations

In the lcSSc subtype, seven non-HLA novel loci were identified as susceptibility markers in the GWAS data (Table S1 and Figure S1). Two out of the seven genetic markers showed evidence of association in the replication cohorts: rs11642873 near the *IRF8* gene (lcSSc $P=2.32\times 10^{-12}$, OR=0.75 [0.69–0.81]) at the GWAS level of significance and rs12540874 in the *GRB10* gene (lcSSc $P=1.27\times 10^{-6}$, OR=1.15 [1.09–1.22]) at the suggestive level of significance (Figure 1, Table 1 and Table S1).

Regarding the dcSSc subtype, five non-HLA loci were found to be associated in the GWAS cohorts (Table S2 and Figure S2). Upon analyzing these five SNPs in the replication cohorts we could only replicate the association of rs11171747 in the *RPLA1/ESYT1* locus (overall dcSSc $P=5.99\times 10^{-8}$, OR=1.23 [1.14–1.33]) (Figure 1, Table 1 and Table S2). However, the association found in this locus was heterogeneous among cohorts (Breslow-Day $P=5.32\times 10^{-9}$).

Auto-Antibodies

The observed associations in the ACA positive subgroup and lcSSc were difficult to differentiate because of substantial overlap between these two disease subgroups. In the GWAS cohorts, SNPs in *IL12RB2* and *RUNX1* genes were identified as novel non-HLA loci associated with SSc patients positive for ACA antibodies (Table S3 and Figure S3). However, none of these associations could be confirmed at the replication stage. Interestingly, the SNP rs11047102 of the *SOX5* gene, which was selected for replication due to its association with the lcSSc subgroup in the GWAS data, showed suggestive evidence of association with the ACA subgroup ($P=1.39\times 10^{-7}$, OR=1.36 [1.21–1.52]) (Figure 1, Table 1 and Table S3).

In the ATA positive subgroup, four new susceptibility loci were identified in the GWAS data (Table S4 and Figure S4), none of which were confirmed in the replication phase. Since the ATA subgroup of patients has the smallest sample size, the lack of replication in any of the non-HLA locus may be due to a lower statistical power (Table S5).

HLA Region

The associations found in the HLA region in the GWAS data set showed clear differences between SSc subgroups (Figure 1, Figure 2, and Table 2). The observed effects in the lcSSc and dcSSc subtype were similar to that of the overlapping group of patients with ACA and ATA respectively, but less significantly. Therefore, we focused the analysis on antibody subgroups only.

We observed independent genetic associations in the ACA positive subgroup in the HLA region (Table 2 and Figure 1, Table S6). The stronger independent signal was identified in the *HLA-DQB1* gene of HLA class II: SNPs rs6457617 (ACA+ $P=1.99\times 10^{-36}$, OR=0.48 [0.42–0.54]) and rs9275390 (ACA+ $P=2.62\times 10^{-54}$, OR=2.38 [2.13–2.67]). The TC allele combination (both risk alleles) showed a high association in the ACA positive subgroup (ACA+ $P=7.81\times 10^{-61}$, OR=2.48 [2.22–2.77]), being present in 45.3% of the ACA positive patients compared to 25.1% of the controls (Table 3).

Regarding the ATA positive subgroup, we also observed evidence of independent association in the HLA region (Table 2 and Figure 1, Table S7). We found three associations in the HLA class II region: rs3129882 in *HLA-DRA* (ATA+ $P=1.89\times 10^{-27}$, OR=2.17 [1.88–2.50]), rs3129763 in the *HLA-DQA1/DRB1* loci (ATA+ $P=1.47\times 10^{-11}$, OR=1.65 [1.42–1.91]) and four associated SNPs in the *HLA-DPA1/DPB1* region (highest association at rs987870, ATA+ $P=2.42\times 10^{-20}$, OR=2.09 [1.78–2.45]). The combination of three risk alleles in the *DPB1/DPB1* locus, CAC (ATA+ $P=1.27\times 10^{-76}$, OR=8.84 [6.72–11.63]) of the SNPs rs987870, rs3135021 and rs6901221 respectively was present in 10.6% of the ATA positive SSc patients compared to only 1.3% of the controls (Table 3).

In addition, in the HLA class III region, the *NOTCH4* gene was associated with the presence of ACA (rs443198, ACA+ $P=8.84\times 10^{-21}$, OR=0.55 [0.49–0.63]) and ATA (rs9296015, ATA+ $P=1.14\times 10^{-8}$, OR=0.54 [0.44–0.67]), independently of the HLA class II associations (Table 2 and Tables S6, S7). Interestingly, SNP rs9296015 had an opposite effect size in ACA and ATA subgroup, being exclusively associated in the ATA subgroup. These two SNPs were not in LD in Caucasian populations either from the HapMap project ($r^2=0.05$ in CEU and $r^2=0.03$ in TSI) or our cohorts ($r^2=0.1$ in the combined cohorts, $r^2=0.11$ in Spanish, $r^2=0.00$ in German, $r^2=0.00$ in Dutch and $r^2=0.01$ in US), pointing to independent associations in the *NOTCH4* gene with both ACA and ATA positive subgroups. All the associations ORs found in the HLA region were consistent among the four GWAS cohorts (Tables S8, S9).

Previously Described Genetic Associations

We wanted to investigate previously reported associations with subphenotypes or overall disease, such as *CD247*, *TNFSF4*, *STAT4*, *BANK1*, *IRF5* and *BLK* in the present study's GWAS cohorts, to further establish them as SSc (or its subphenotypes) susceptibility loci. Table S10 shows the analysis of the SNPs in the previously mentioned genes which were present in our GWAS combined panel. As expected, association previously found in these six genes was replicated. Interestingly associations previously described to be confined to one of the SSc subgroups were also replicated as in the cases of *TNFSF4* and lcSSc (lcSSc $P=7.70\times 10^{-4}$, OR=1.18 [1.03–1.31]), *STAT4* and lcSSc (lcSSc $P=7.70\times 10^{-8}$, OR=1.31 [1.19–1.48]), *BANK1* and dcSSc (dcSSc $P=0.0103$, OR=0.85 [0.75–0.96]) and *BLK* and ACA+ (ACA+ $P=1.45\times 10^{-4}$, OR=1.27 [1.12–1.44]). Furthermore association of *CD247* with SSc was more strongly represented in the lcSSc subgroup than the others (lcSSc $P=2.66\times 10^{-6}$, OR=0.81 [0.75–0.89]), although evidence of association was also

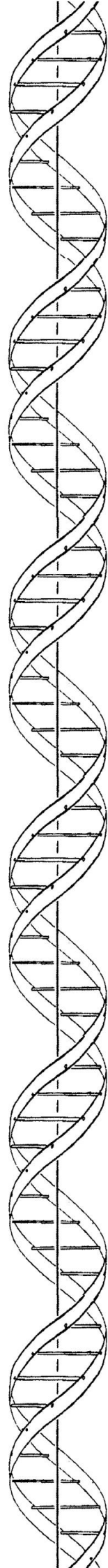


Table 1. Novel non-HLA loci associated with SSc clinical and serological subtypes.

SSc Subphenotype	Chr.	Gene	SNP	Base Pair	Location	Change	Stage	N (case/control)	MAF (case/control)	P† value	OR (95% CI)
lcSSc	7p12.1	GRB10	rs12540874	50,632,416	Intronic	G/A	GWAS	1400/5172	0.461/0.409	3.00×10^{-6}	1.23 (1.13-1.34)
							Replication	1960/4971	0.416/0.395	3.07×10^{-2}	1.09 (1.01-1.18)
							Combined	3360/10143	0.435/0.403	1.27×10^{-6}	1.15 (1.09-1.22)
	16q24.1	IRF8	rs11642873	84,549,206	Intergenic	C/A	GWAS	1400/5172	0.144/0.197	1.39×10^{-7}	0.72 (0.64-0.81)
							Replication	1960/4971	0.143/0.186	6.88×10^{-6}	0.78 (0.70-0.87)
							Combined	3360/10143	0.144/0.192	2.32×10^{-12}	0.75 (0.69-0.81)
dcSSc	12q13.2	RPL41/ESYT1*	rs11171747	54,804,675	Upstream	G/T	GWAS	740/5172	0.446/0.384	2.19×10^{-6}	1.31 (1.01-1.29)
							Replication	959/4971	0.408/0.372	3.49×10^{-3}	1.16 (1.15-1.71)
							Combined	1699/10143	0.425/0.379	5.99×10^{-8}	1.23 (1.10-1.50)
ACA+	12p12.1	SOX5	rs11047102	23,837,413	Intronic	T/C	GWAS	761/5172	0.132/0.096	1.03×10^{-5}	1.47 (1.24-1.73)
							Replication	1030/4971	0.123/0.102	2.91×10^{-3}	1.27 (1.09-1.48)
							Combined	1791/10143	0.127/0.099	1.39×10^{-7}	1.36 (1.21-1.52)

†P values for GWAS cohorts are Mantel-Haenszel meta-analysis GC corrected according to the set λ and in the replication and combined analysis Mantel-Haenszel meta-analysis P value.

*Association in rs11171747 had a significant BD P value, thus making it heterogeneous association among populations.

doi:10.1371/journal.pgen.1002178.t001

found in the other subgroups. Similarly, the association found in *IRF5* was stronger in lcSSc (lcSSc $P=1.64 \times 10^{-10}$, OR = 1.50 [1.32–1.69]), although association was also found in the dcSSc, ACA+ and ATA+ subgroups.

Discussion

Systemic sclerosis (SSc) is a rare, severe, complex and heterogeneous rheumatic disease. Multiple lines of evidence

suggest that genetic factors may underlie not only SSc susceptibility but also the predisposition to develop specific clinical phenotypes such as lcSSc, dcSSc subtypes and the presence of SSc-specific auto-antibodies. The discovery of genetic variants associated with specific clinical manifestations of the disease will lead to new insights regarding pathogenesis and may open novel avenues of therapy that can be targeted to specific subsets.

The aim of this study was to assess the genetic component involved in four different SSc clinical and auto-antibody

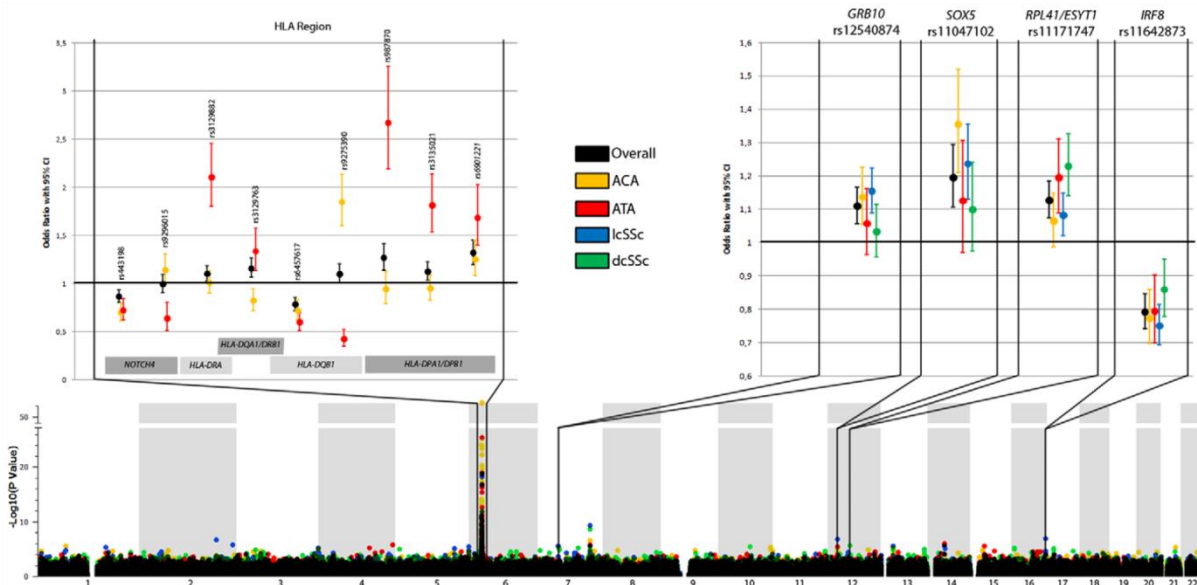


Figure 1. New loci associated with subphenotypes of SSc. The lower part shows the Manhattan Plot with corrected P values of the GWAS cohorts. The upper part shows the ORs and the 95% CI interval of the novel associated regions in the GWAS cohorts (HLA region, left panel) and all cohorts (non-HLA loci, right panel) for the overall analysis and each subphenotype considered in the study. (Note: the ORs and CIs on the forest plot do not exactly correspond to the numbers in Table 1 and Table 2. Table 1 and Table 2 shows marginal effects of these SNPs while this figure presents ORs and CIs after the adjustment for the other SNPs claimed as independent for that phenotype).

doi:10.1371/journal.pgen.1002178.g001

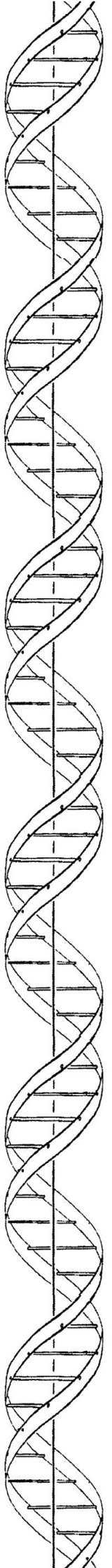
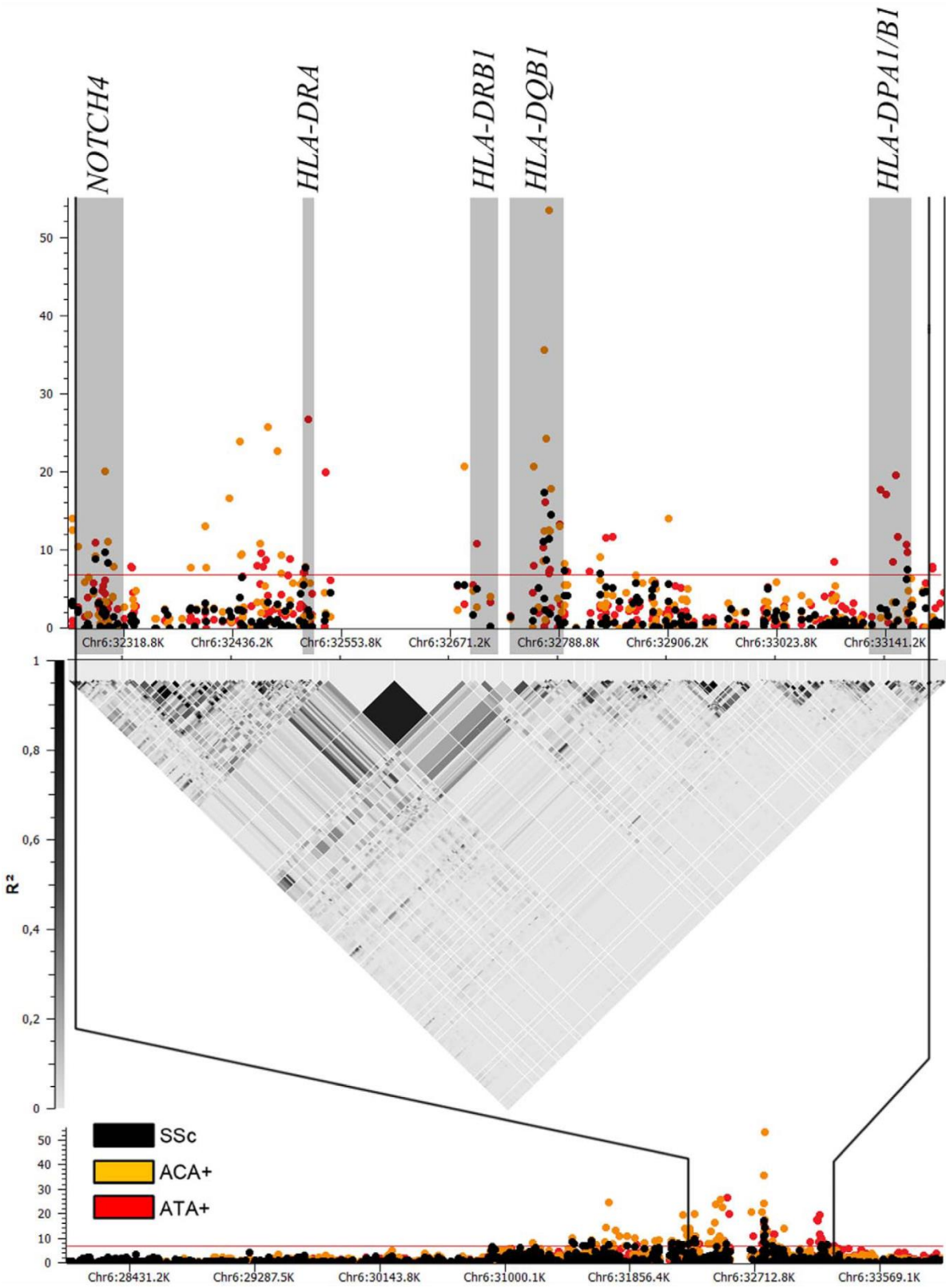


Figure 2. Manhattan plot showing the $-\log_{10}$ of the Mantel-Haenszel P value of all 1,112 SNPs in HLA region for the GWAS cohorts comprising 2,296 cases and 5,171 controls. Associations for the whole SSc set are in black, while associations in ACA (760 cases) and ATA (447 cases) positive subgroups are in orange and red, respectively. Loci which were independently associated according to conditional logistic regression analysis are highlighted in grey.

doi:10.1371/journal.pgen.1002178.g002

subphenotypes through an analysis of our previous genome-wide association study (GWAS) data stratified for these disease subphenotypes, together with a large, new replication study.

We have identified an association of the *NOTCH4* gene with both ACA and ATA positive subgroups independent of the HLA associations. This gene is located in the MHC and encodes a transmembrane protein which plays a role in a variety of developmental processes by controlling cell fate decisions. Interestingly, *NOTCH4* has been implicated in the pathways by which TGF- β induces pulmonary fibrosis [24], one of the most severe clinical manifestations of SSc [25,26]. The Notch signaling pathway also controls key functions in vascular smooth muscle and endothelial cells which may be particularly relevant to the microvascular damage seen in SSc [27]. Genetic variants in *NOTCH4* also have been previously associated, independently from HLA genes or alleles, with other autoimmune disorders like diabetes type 1 [28], rheumatoid arthritis [29] and alopecia areata [30,31].

Additionally, through the analysis of the largest SSc case/control cohort reported to date we identified three new susceptibility loci (*IRF8*, *SOX5* and *GRB10*), outside the HLA/MHC region, implicated in genetic predisposition to different SSc subphenotypes, in addition to other suggestive loci.

Type I and II interferons (IFN) are well known immunomodulators which can also regulate collagen production. Furthermore, they are believed to play a key role in the pathogenesis of SSc and other autoimmune diseases [32–34]. Interestingly, we found a strong association of the *IRF8* gene with the lcSSc subtype and the ACA positive subgroup. *IRF8* modulates TLR signaling and may contribute to the crosstalk between IFN- γ and TLR signal pathways, thus acting as a link between innate and adaptive immune responses [35]. *IRF8* also has been demonstrated to be a key factor in B cell lineage specification, commitment and differentiation [36]. In addition, *IRF8* has been associated with another autoimmune disease, multiple sclerosis [37], although the SNP associated with multiple sclerosis (rs17445836) was not present in our study. Nevertheless, both variants are in medium LD in the CEU population of the HapMap project ($r^2 = 0.51$) and both associations have a protective OR for the minor allele; pointing to a dependence in the associations found in these two diseases.

The most prominent SSc specific auto-antibodies, ACA and ATA, are associated with the lcSSc and dcSSc clinical subsets, respectively [19]. The lcSSc subtype greatly overlaps with the ACA positive subgroup of patients (almost all ACA positive patients belonged to the lcSSc subtype). Similarly, the dcSSc subtype overlaps with the ATA positive group of patients. Therefore, it is difficult to determine whether some of the observed associations specifically belonged to one of the four subgroups. Such is the case of the association found with the *SOX5* gene. In the GWAS data, *SOX5* was associated with lcSSc as well as with the ACA positive subgroup, although the association with the lcSSc subtype was stronger than that in the ACA positive subgroup. Upon completion of the replication study with the resultant increase in statistical power, we were able to determine that the *SOX5* gene was indeed a risk factor for the ACA positive group at the genome wide significance level, but not for lcSSc. The *SOX5* gene encodes a member of the SOX (*SRY*-related HMG-

box) family of transcription factors involved in the regulation of embryonic development, in the determination of cell fate, as well as in chondrogenesis [38].

Conversely *SOX5*, together with *SOX6* and *SOX9*, can induce many cellular types (including melanocytes and bone marrow stem cells) into the chondrogenic pathway, leading to expression of *COL2A1* and the formation of cartilage [38,39]. As stated above, IFN type I and II are inhibitors of collagen production and chondrogenesis; more precisely IFN- γ (type II IFN) inhibits the *COL2A1* gene which is one of the main downstream genes in the chondrogenesis pathway [40]. Taken all together, *IRF8* (part of the interferon pathway and induced by IFN- γ [41]) and *SOX5* may be affecting the formation of the extra-cellular matrix through *COL2A1* in the skin and other organs of SSc patients.

We also identified an association of the *GRB10* gene with the lcSSc subtype; *GRB10* codes for an adaptor protein known to interact with a number of tyrosine kinase receptors and signaling molecules and has a potential role in apoptosis regulation [42].

In dcSSc patients, the only observed genome wide significant association was with the *RPLA1/ESYTT1* locus, although this association was heterogeneous among the investigated populations, probably due to lower statistical power in this smaller group. Three genes are relevant to this locus: *RPLA1*, a ribosomal protein not considered to be related to the immune system; *ZC3H10*, a zinc finger protein related to tumour growth; and *ESYTT1*, a synaptotagmin-like protein of unknown function. Although none of these genes has a suggestive role in the pathogenesis of SSc *a priori*, further studies are needed to investigate this intriguing finding.

Since most genes in the HLA region are implicated in the regulation of the immune system, it is not surprising that the HLA-association with SSc is primarily related to auto-antibody expression. We found different patterns of independent association for the two major SSc auto-antibody subgroups across the HLA class II region. Both genetic markers located in the *HLA-DQB1* locus were associated with the presence of ACA auto-antibodies in SSc patients. The allelic combination of these SNPs tags the described association of HLA-DQB1*0501 with the ACA positive subgroup of the disease [22,43]. The associations within the HLA region in the ATA positive subgroup are more complex: SNP rs3129763 (located near *HLA-DRB1*) tags the association of HLA-DRB1*1104, which has been described to be associated with the whole disease [22]. Furthermore, the haplotype in the *HLA-DPB1* region described in Table 3, tags the HLA-DPB1*1301 also previously described [3,22]. Interestingly, the remaining independent association observed, rs3129882, is found within the *HLA-DRA* gene, which is much less polymorphic than the other HLA genes already mentioned; nevertheless, the association found in this SNP is tagging through the extensive LD structure of the MHC region the association of some aminoacidic positions in the nearby *HLA-DQB1* gene, which has not been previously reported to be associated with the ATA positive subgroup of SSc.

In summary, taking advantage of our GWAS data and a large replication cohort, we have identified three new non-HLA loci associated with subphenotypes of SSc: *GRB10*, *IRF8*, and *SOX5*. In addition, we shed light on HLA associations with this disease, establishing different patterns of independent association in the ACA and ATA positive subgroups. Our findings provide evidence for genetic heterogeneity underlying the clinical and especially

Table 2. Independent associations identified in the HLA region with the ACA and ATA positive subgroups.

SSc Subphenotype	SNP	Gene	Location	Change	MAF (ACA/ATA/control)	ACA			ATA				
						Unadjusted		Adjusted		Unadjusted		Adjusted	
						P^{\dagger}	OR (CI 95%)	P^*	OR (CI 95%)	P^{\dagger}	OR (CI 95%)	P^*	OR (CI 95%)
ACA+	rs443198	NOTCH4	Exon	C/T	0.253/0.304/0.371	8.83×10^{-21}	0.55 (0.49–0.63)	7.412×10^{-8}	0.70 (0.09–0.10)	3.91×10^{-5}	3.89×10^{-5}	0.72 (0.10–0.12)	
	rs6457617	HLA-DQB1	Intergenic	C/T	0.314/0.442/0.492	1.99×10^{-36}	0.48 (0.42–0.54)	1.67×10^{-5}	0.72 (0.10–0.12)	0.00427	2.68×10^{-10}	0.60 (0.09–0.10)	
	rs9275390	HLA-DQB1	Intergenic	C/T	0.454/0.177/0.253	2.61×10^{-54}	2.38 (2.13–2.67)	4.793×10^{-17}	1.85 (0.25–0.29)	9.70×10^{-8}	4.45×10^{-16}	0.43 (0.08–0.10)	
ATA+	rs9296015	NOTCH4	Intergenic	A/G	0.214/0.117/0.186	0.1161	1.11 (0.97–1.27)	0.0611	1.14 (0.15–0.17)	1.14×10^{-8}	0.000122	0.64 (0.13–0.16)	
	rs3129882	HLA-DRA	Intron	G/A	0.430/0.631/0.440	0.2725	0.94 (0.84–1.05)	0.867	1.01 (0.11–0.12)	1.893×10^{-27}	4.58×10^{-21}	2.11 (0.30–0.35)	
	rs3129763	HLA-DQA1/DRB1	Intergenic	A/G	0.209/0.348/0.246	0.00221	0.81 (0.71–0.93)	0.00687	0.82 (0.11–0.12)	1.474×10^{-11}	0.000518	1.34 (0.20–0.24)	
	rs987870	HLA-DPA1/DPB1	Intron	C/T	0.139/0.270/0.146	0.1725	0.89 (0.76–1.05)	0.525	0.94 (0.15–0.18)	2.419×10^{-20}	1.40×10^{-22}	2.67 (0.48–0.58)	
	rs3135021	HLA-DPA1/DPB1	Intron	A/G	0.271/0.403/0.286	0.0839	0.90 (0.79–1.01)	0.463	0.95 (0.12–0.14)	1.949×10^{-12}	2.02×10^{-12}	1.81 (0.28–0.33)	
	rs6901221	HLA-DPA1/DPB1	Intron	C/A	0.190/0.223/0.157	2.98×10^{-5}	1.35 (1.17–1.55)	0.00252	1.25 (0.17–0.20)	2.542×10^{-8}	2.55×10^{-8}	1.69 (0.28–0.34)	

Sample size for the ACA subgroup was 761 and for ATA was 447, while the sample size for the controls was 5,172.

[†]Unadjusted P values are Mantel-Haenszel meta-analysis, GC corrected for the λ of the set, of all GWAS cohorts.

*Adjusted P values are logistic regression analysis adjusted for all other SNPs in the same region and the same subphenotype. doi:10.1371/journal.pgen.1002178.t002

autoantibody subtypes of SSc. These findings may prompt reconsideration of the current classification of SSc patients; provide insight into pathogenetic pathways differing among subphenotypes, especially specific auto-antibody subgroups, and lead to novel therapeutic targets for this devastating autoimmune disease.

Materials and Methods

Subjects

For the GWAS analysis, a total of 2,296 Caucasian SSc patients and 5,171 Caucasian healthy controls were recruited through an international collaborative effort in the United States of America (USA), Spain, Germany and The Netherlands. The North American cases (initial $n=1,678$; after applying quality control criteria, $n=1,486$; 179 men, 1,307 women; mean age = 54.5 (median, 55.0); SD = 12.9) were recruited from May, 2001 to December, 2008 from three U.S. sources: the Scleroderma Family Registry and DNA Repository and the Center of Research Translation in Scleroderma at The University of Texas (UT) Health Science Center-Houston, The Johns Hopkins University Medical Center and the Fred Hutchinson Cancer Research Center, each enrolling patients from a US-wide catchment area. The initial European SSc cases came from previously established nationally representative collections of 380 Spanish, 288 German and 190 Dutch patients with SSc. As control populations, healthy unrelated individuals of Spanish (initial $n=414$), German (initial $n=678$) and Dutch (initial $n=643$) origin were included in the study as well as 3478 controls from across the US collected as non-cancer controls for GWAS studies of breast and prostate cancers in the Cancer Genetic Markers of Susceptibility (CGEMS) studies [44,45] (<http://cgems.cancer.gov/data/>).

In the second replication phase, a large independent replication cohort, consisting of 3,175 SSc patients and 4,971 healthy controls of Caucasian ancestry, were collected from Belgium, Spain, The Netherlands, Germany, Italy, Norway, Sweden, UK and the USA. Details on the investigated populations are provided in the Table S11.

All cases met the American College of Rheumatology preliminary criteria for the classification of SSc [46]. Furthermore, patients were classified according to the extent of skin involvement into limited (lcSSc) or diffuse (dcSSc) forms [17,47]. In addition, the presence of SSc specific auto-antibodies, anti-topoisomerase I (ATA, Anti-Scl70) and anti-centromere (ACA) was assessed by passive immunodiffusion against calf thymus extract (Inova Diagnostics, San Diego, CA, USA) and indirect immunofluorescence of HEp-2 cells (Antibodies Inc, Davis, CA, USA), respectively, in a total of 5,229 and 5,238 SSc patients respectively. Auto-antibodies to RNA Polymerase III are also considered to be characteristic of SSc, but testing for this antibody is not widely available and since results were not known in almost two-thirds of our cases, this analysis was not done [18,19]. The distribution of SSc patients among these disease subsets is summarized in Table S11.

Collection of blood samples and clinical information from case and control subjects was undertaken with informed consent and relevant ethical review board approval from each contributing centre in accordance with the tenets of the Declaration of Helsinki.

Most of the individuals included in this study, GWAS and replication cohorts, have been analyzed in a previous study [15] but novel genotypes were generated in the replication cohorts for phenotype associated SNPs found in the GWAS, expanding the scope of the study.

SNP Selection for Replication

Our goal was to examine any novel genetic association specific for each subset rather than overall disease. Although partial

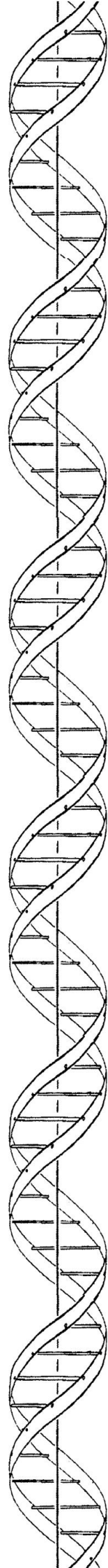


Table 3. Allelic combination analysis of the SNPs which are in the same association locus within the HLA region for the ACA and ATA positive subgroups of SSc patients.

SSc Subphenotype	Locus	Haplotype	N (case/control)	Frequency (case/control)	P Value	OR (CI 95%)	SNPs
ACA	HLA-DQB1	TC	761/5172	0.453/0.251	7.807×10^{-61}	2.48 (2.22–2.77)	rs6457617 rs9275390
		CT	761/5172	0.313/0.490	3.639×10^{-38}	0.47 (0.42–0.53)	rs6457617 rs9275390
		TT	761/5172	0.234/0.259	0.0353	0.87 (0.77–0.99)	rs6457617 rs9275390
ATA	HLA-DP	CAC	447/5172	0.106/0.013	1.266×10^{-76}	8.84 (6.72–11.63)	rs987870 rs3135021 rs6901221
		TAC	447/5172	0.019/0.012	0.0745	1.55 (0.92–2.60)	rs987870 rs3135021 rs6901221
		TGC	447/5172	0.101/0.132	0.00792	0.74 (0.59–0.92)	rs987870 rs3135021 rs6901221
		TAA	447/5172	0.265/0.256	0.562	1.05 (0.90–1.23)	rs987870 rs3135021 rs6901221
		CGA	447/5172	0.148/0.127	0.0798	1.20 (0.98–1.46)	rs987870 rs3135021 rs6901221
		TGA	447/5172	0.361/0.460	2.137×10^{-8}	0.67 (0.58–0.77)	rs987870 rs3135021 rs6901221

doi:10.1371/journal.pgen.1002178.t003

overlapping exists between lcSSc and ACA+ subgroups, and dcSSc and ATA+ subgroups; we wanted to assess whether association found in overlapped groups belonged to a subtype or an auto-antibody positive group. With that purpose we selected SNPs from the GWAS data based on the following criteria:

- First, we selected all SNPs with a P value of 1×10^{-5} or lower in each of the four considered SSc subgroups (*i.e.* lcSSc, dcSSc, ACA+ and ATA+) of the four GWAS cohorts (*i.e.* US, Spain, Netherlands and Germany).
- Since one aim of this study was to find novel genetic associations, we then ruled out every genetic association previously described in SSc (*e.g.* *STAT4*, *IRF5* and the HLA region).
- To select subphenotype specific signals, we excluded all SNPs with P values of the same order of magnitude or lower in the opposite group, *i.e.* lcSSc versus dcSSc and ACA-positive versus ATA-positive.
- Finally we selected from each remaining region the best independent association (determined by conditional logistic regression) from the GWAS data.

This resulted in the selection of 18 non-HLA SNPs (7 for lcSSc, 5 for dcSSc, 2 for ACA+, and 4 for ATA+) as shown in Tables S1, S2, S3, S4, corresponding to lcSSc, dcSSc, ACA and ATA positive patients respectively.

Genotyping

The GWAS genotyping of the SSc cases and controls was performed as follows: the Spanish SSc cases and controls together with Dutch and German SSc cases was performed at the Department of Medical Genetics of the University Medical Center Utrecht (The Netherlands) using the commercial release Illumina HumanCNV370K BeadChip, which contains 300,000 standard SNPs with an additional 52,167 markers designed to specifically target nearly 14,000 copy number variant regions of the genome, for a total of over 370,000 markers. Genotype data for Dutch and German controls were obtained from the Illumina Human 550K BeadChip available from a previous study. The SSc case group from the United States was genotyped at Boas Center for Genomics and Human Genetics, Feinstein Institute for Medical Research, North Shore Long Island Jewish Health System using the Illumina Human610-Quad BeadChip. CGEMS and Illumina iControlDB controls were genotyped on the Illumina Hap550K-BeadChip.

SNPs selected for the replication phase were genotyped in the replication cohorts using Applied Biosystems' TaqMan SNP assays on ABI Prism 7900 HT real-time thermocyclers. Markers with call rates of 95% or less were excluded, as were markers whose allele distributions deviated strongly from Hardy-Weinberg (HW) equilibrium in controls ($P < 10^{-3}$).

Data Imputation

Imputation was performed in the GWAS cohorts in order to gain genome coverage for the SNP selection. Imputation was performed with IMPUTE software 1.00 as previously described [48], using as reference panels the CEU and TSI HapMap populations. However, SNP imputation did not show any new independent SNP associated at $P < 10^{-5}$ in the four subphenotypes considered. The imputed GWAS data in the four subphenotypes is shown in Figure S5.

Statistical Analysis

Data in the SSc GWAS cohorts was filtered as follows: Using Plink, we identified and excluded pairs of genetically related subjects or duplicates and excluded the genetic-pair members with lower call rates. To identify individuals who might have non-western European ancestry, we merged our case and control data with the data from the HapMap Project (60 western European (CEU), 60 Nigerian (YRI), 90 Japanese (JPT) and 90 Han Chinese (CHB) samples). We used principal component analysis as implemented in HelixTree (see Text S2), plotting the first two principal components for each individual. All individuals who did not cluster with the main CEU cluster (defined as deviating more than 4 standard deviations from the cluster centroids) were excluded from subsequent analyses. Additionally, we excluded individuals with low call rates (11 individuals from the US group, 24 from the Spanish, 1 from the German and 1 from the Dutch), relatedness (50 from the US group, 2 from the Spanish, 1 from the German and 1 from the Dutch), non-European ancestry (42 from the US group, 5 from the Spanish, 6 from the German and 4 from the Dutch) and inconsistent gender (83 from the US group, 2 from the Spanish, 2 from the German and 2 from the Dutch). Then we filtered for SNP quality, removing SNPs with a genotyping success call rate $< 98\%$ and those showing $MAF < 1\%$. Deviation of the genotype frequencies in the controls from those expected under Hardy-Weinberg equilibrium was assessed by a χ^2 test or Fisher's exact test when an expected cell count was < 5 . SNPs strongly deviating from Hardy-Weinberg equilibrium ($P < 10^{-5}$) were

eliminated from the study. For the combined analysis of the four datasets, the same quality controls per individual and per SNP were applied with the exception of the Hardy-Weinberg equilibrium (HWE) requirement. The genotyping success call rate on the merged dataset after all these quality filters were applied was 99.83% in the GWAS cohorts.

The replication cohorts were filtered as follows: all individuals with a SNP success call rate below 0.95 were excluded, SNPs with a per individual success call rate below 0.95 were excluded, SNPs with a HWE comparison P value below 0.001 in controls were excluded and SNPs with a MAF below 0.01 were also excluded. As a result, 18 SNPs selected for replication all were in HWE (P value > 0.001) and the overall genotype successful call rate was 96.61% and all SNPs individually had a successful call rate greater than 95%.

We performed power calculations for GWAS and replication cohorts for the whole dataset and the clinical/auto-antibodies subphenotypes according to Skol *et al.* [49] (Table S5). The significance level for these calculations was set at 5×10^{-8} .

χ^2 tests were performed for allelic model for significant differences between cases and controls. Derived P values for the replication cohorts were not adjusted. All nine replication cohorts were jointly analyzed conducting Cochran-Mantel-Haenszel (CMH) tests to control for population differences. A threshold meta-analysis P value of < 0.05 for the replication phase was considered significant. We also conducted CMH meta-analysis of all the nine replication cohorts and the four cohorts previously included in the GWAS, considering a P value lower than 5×10^{-8} as significant. Furthermore, P values in the range 5×10^{-8} to 5×10^{-6} were considered as suggestive associations. In all tests, odds ratios (OR) were calculated according to Woolf's method. We also applied Breslow-Day (BD) tests for all meta-analyses to check for heterogeneity in association among the investigated populations, and all associations with a $P < 0.05$ in BD analysis were considered heterogeneous.

Due to the partial overlapping of the lcSSc and dcSSc subgroups with ACA+ and ATA+ subgroups, respectively, we wanted to test whether an association found in both overlapping groups belonged to one or the other specifically. With that purpose, all the associations in the present study claimed to belong to a group were tested for association in the correlated group (*e.g.* ACA associations were tested in lcSSc and vice versa) to look for the best P value. In addition, ACA and ATA hits were tested in lcSSc-ACA- and dcSSc-ATA-, respectively, to ensure group specific associations. Also, lcSSc and dcSSc were tested in ACA+-non-lcSSc and ATA+-non-dcSSc with the same purpose.

To determine independent associations in the HLA region, conditional logistic regression was carried out for all associated SNPs in the complete SSc group and the ACA and ATA positive subgroups. This analysis was carried out as implemented in Plink software, conditioning each SNP association to each of the other significantly associated ($P < 5 \times 10^{-7}$) SNPs in the corresponding LD block, controlling for the presence of the four populations as covariates. All SNPs which remained significant after conditioning were considered independent associations. All haplotype analysis was performed using Haploview software, defining the blocks by confidence intervals [50]. We only analyzed haplotypes or allelic combinations with frequencies of 1% and above.

Statistical analyses were undertaken using R (v2.6), Stata (v8), Plink (v1.07) [51] and HelixTree's SNP & Variation Suite (v7.3.0) software (see Text S2).

Web Resources

Plink software:

<http://pngu.mgh.harvard.edu/purcell/plink/>

SVS HelixTree software:

http://www.goldenhelix.com/SNP_Variation/HelixTree/index.html

Stata software:

<http://www.stata.com/>

R Statistical Package:

<http://www.r-project.org/>

Haploview:

<http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/haploview>

Supporting Information

Figure S1 Manhattan plot and QQ plot showing the $-\log_{10}$ of the Mantel-Haenszel P value of all 279,621 SNPs in the lcSSc individuals of the GWAS cohorts comprising 1,400 cases and 5,171 controls. All P values are GC corrected, and λ was 1.058. (TIF)

Figure S2 Manhattan plot and QQ plot showing the $-\log_{10}$ of the Mantel-Haenszel P value of all 279,621 SNPs in the dcSSc individuals of the GWAS cohorts comprising 740 cases and 5,171 controls. All P values are GC corrected, and λ was 1.034. (TIF)

Figure S3 Manhattan plot and QQ plot showing the $-\log_{10}$ of the Mantel-Haenszel P value of all 279,621 SNPs in the ACA positive individuals of the GWAS cohorts comprising 761 cases and 5,171 controls. All P values are GC corrected, and λ was 1.050. (TIF)

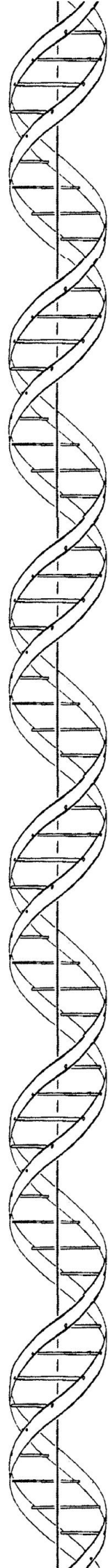
Figure S4 Manhattan plot and QQ plot showing the $-\log_{10}$ of the Mantel-Haenszel P value of all 279,621 SNPs in the ATA positive individuals of the GWAS cohorts comprising 447 cases and 5,171 controls. All P values are GC corrected, and λ was 1.061. (TIF)

Figure S5 Manhattan plot showing the analysis in the GWAS cohorts imputed data. The different subphenotypes considered are represented in different colors. (TIF)

Table S1 Analysis for GWAS cohorts, replication cohorts and combined analysis for all non-HLA, non-previously described associations with lcSSc subtype of the disease. † P values for GWAS cohorts are Mantel-Haenszel meta-analysis GC corrected according to the set λ and in the replication and combined analysis Mantel-Haenszel meta-analysis P value. ‡ P value for the totality of the SSc patients, in the case of GWAS cohorts GC corrected according to the set λ , and in replication and combined analysis Mantel-Haenszel meta-analysis P value. (DOC)

Table S2 Analysis for GWAS cohorts, replication cohorts and combined analysis for all non-HLA, non-previously described associations with dcSSc subtype of the disease. † P values for GWAS cohorts are Mantel-Haenszel meta-analysis GC corrected according to the set λ and in the replication and combined analysis Mantel-Haenszel meta-analysis P value. ‡ P value for the totality of the SSc patients, in the case of GWAS cohorts GC corrected according to the set λ , and in replication and combined analysis Mantel-Haenszel meta-analysis P value. *Association in rs11171747 had a significant BD P value, thus making them heterogenic associations among populations. (DOC)

Table S3 Analysis for GWAS cohorts, replication cohorts and combined analysis for all non-HLA, non-previously described



associations with ACA positive subgroup of the disease. †*P* values for GWAS cohorts are Mantel-Haenszel meta-analysis GC corrected according to the set λ and in the replication and combined analysis Mantel-Haenszel meta-analysis *P* value. ‡*P* value for the totality of the SSc patients, in the case of GWAS cohorts GC corrected according to the set λ , and in replication and combined analysis Mantel-Haenszel meta-analysis *P* value. *Association in rs3790567 had a significant BD *P* value, thus making them heterogeneous associations among populations. (DOC)

Table S4 Analysis for GWAS cohorts, replication cohorts and combined analysis for all non-HLA, non-previously described associations with ATA positive subgroup of the disease. †*P* values for GWAS cohorts are Mantel-Haenszel meta-analysis GC corrected according to the set λ and in the replication and combined analysis Mantel-Haenszel meta-analysis *P* value. ‡*P* value for the totality of the SSc patients, in the case of GWAS cohorts GC corrected according to the set λ , and in replication and combined analysis Mantel-Haenszel meta-analysis *P* value. (DOC)

Table S5 Power calculations and genomic inflation factors (λ) in the whole SSc cohorts (GWAS and replication) and the lcSSc, dcSSc, ACA and ATA positive subphenotypes. 5×10^{-8} was used as significance threshold. (DOC)

Table S6 Conditional logistic regression analysis of all the independently associated SNPs in the HLA region in the ACA positive patients. †*P* values for Mantel-Haenszel meta-analysis GC corrected according to the set λ . (DOC)

Table S7 Conditional logistic regression analysis of all the independently associated SNPs in the HLA region in the ATA positive patients. †*P* values for Mantel-Haenszel meta-analysis GC corrected according to the set λ . (DOC)

Table S8 Independent associations found in the HLA region in the ACA positive subgroup of patients in the separate four GWAS cohorts. †Uncorrected χ^2 *P* value of each separated cohort. (DOC)

References

- Agarwal SK, Tan FK, Arnett FC (2008) Genetics and genomic studies in scleroderma (systemic sclerosis). *Rheum Dis Clin North Am* 34: 17–40.
- Arnett FC, Gourh P, Shete S, Ahn CW, Honey RE, et al. (2010) Major histocompatibility complex (MHC) class II alleles, haplotypes and epitopes which confer susceptibility or protection in systemic sclerosis: analyses in 1300 Caucasian, African-American and Hispanic cases and 1000 controls. *Ann Rheum Dis* 69: 822–827.
- Zhou X, Lee JE, Arnett FC, Xiong M, Park MY, et al. (2009) HLA-DPB1 and DPB2 are genetic loci for systemic sclerosis: a genome-wide association study in Koreans with replication in North Americans. *Arthritis Rheum* 60: 3807–3814.
- Rueda B, Broen J, Simeon C, Hesselstrand R, Diaz B, et al. (2009) The STAT4 gene influences the genetic predisposition to systemic sclerosis phenotype. *Hum Mol Genet* 18: 2071–2077.
- Dieude P, Guedj M, Wipff J, Ruiz B, Hachulla E, et al. (2009) STAT4 is a genetic risk factor for systemic sclerosis having additive effects with IRF5 on disease susceptibility and related pulmonary fibrosis. *Arthritis Rheum* 60: 2472–2479.
- Tsuchiya N, Kawasaki A, Hasegawa M, Fujimoto M, Takehara K, et al. (2009) Association of STAT4 polymorphism with systemic sclerosis in a Japanese population. *Ann Rheum Dis* 68: 1375–1376.
- Dieude P, Guedj M, Wipff J, Avouac J, Fajardy I, et al. (2009) Association between the IRF5 rs2004640 functional polymorphism and systemic sclerosis: a new perspective for pulmonary fibrosis. *Arthritis Rheum* 60: 225–233.
- Ito I, Kawaguchi Y, Kawasaki A, Hasegawa M, Ohashi J, et al. (2009) Association of a functional polymorphism in the IRF5 region with systemic sclerosis in a Japanese population. *Arthritis Rheum* 60: 1845–1850.
- Gourh P, Agarwal SK, Martin E, Divecha D, Rueda B, et al. (2010) Association of the C8orf13-BLK region with systemic sclerosis in North-American and European populations. *J Autoimmun* 34: 155–162.
- Ito I, Kawaguchi Y, Kawasaki A, Hasegawa M, Ohashi J, et al. (2010) Association of the FAM167A-BLK region with systemic sclerosis. *Arthritis Rheum* 62: 890–895.
- Dieude P, Wipff J, Guedj M, Ruiz B, Melchers I, et al. (2009) BANK1 is a genetic risk factor for diffuse cutaneous systemic sclerosis and has additive effects with IRF5 and STAT4. *Arthritis Rheum* 60: 3447–3454.
- Rueda B, Gourh P, Broen J, Agarwal SK, Simeon C, et al. (2010) BANK1 functional variants are associated with susceptibility to diffuse systemic sclerosis in Caucasians. *Ann Rheum Dis* 69: 700–705.
- Gourh P, Arnett FC, Tan FK, Assassi S, Divecha D, et al. (2010) Association of TNFSF4 (OX40L) polymorphisms with susceptibility to systemic sclerosis. *Ann Rheum Dis* 69: 550–555.
- Bossini-Castillo L, Broen JC, Simeon CP, Beretta L, Vonk MC, et al. (2011) A replication study confirms the association of TNFSF4 (OX40L) polymorphisms with systemic sclerosis in a large European cohort. *Ann Rheum Dis* 70: 638–641.
- Radstake TR, Gorlova O, Rueda B, Martin JE, Alizadeh BZ, et al. (2010) Genome-wide association study of systemic sclerosis identifies CD247 as a new susceptibility locus. *Nat Genet* 42: 426–429.
- Jimenez SA, Derk CT (2004) Following the molecular pathways toward an understanding of the pathogenesis of systemic sclerosis. *Ann Intern Med* 140: 37–50.

Systemic Sclerosis Novel Genetic Associations

17. LeRoy EC, Medsger TA, Jr. (2001) Criteria for the classification of early systemic sclerosis. *J Rheumatol* 28: 1573–1576.
18. Gabrielli A, Avvedimento EV, Krieg T (2009) Scleroderma. *N Engl J Med* 360: 1989–2003.
19. Steen VD (2008) The many faces of scleroderma. *Rheum Dis Clin North Am* 34: 1–15.
20. Nierert PJ, Mitchell HC, Bolster MB, Shaftman SR, Tilley BC, et al. (2006) Racial variation in clinical and immunological manifestations of systemic sclerosis. *J Rheumatol* 33: 263–268.
21. Assassi S, Arnett FC, Reveille JD, Gourh P, Mayes MD (2007) Clinical, immunologic, and genetic features of familial systemic sclerosis. *Arthritis Rheum* 56: 2031–2037.
22. Arnett FC, Gourh P, Shete S, Ahn CW, Honey R, et al. (2009) Major Histocompatibility Complex (MHC) class II alleles, haplotypes, and epitopes which confer susceptibility or protection in the fibrosing autoimmune disease systemic sclerosis: analyses in 1300 Caucasian, African-American and Hispanic cases and 1000 controls. *Ann Rheum Dis* 69(5): 822–7.
23. Gourh P, Tan FK, Assassi S, Ahn CW, McNearney TA, et al. (2006) Association of the PTPN22 R620W polymorphism with anti-topoisomerase I- and anticentromere antibody-positive systemic sclerosis. *Arthritis Rheum* 54: 3945–3953.
24. Hardie WD, Korfhagen TR, Sartor MA, Prestridge A, Medvedovic M, et al. (2007) Genomic profile of matrix and vasculature remodeling in TGF- α induced pulmonary fibrosis. *Am J Respir Cell Mol Biol* 37: 309–321.
25. Silver RM, Miller KS, Kinsella MB, Smith EA, Schabel SI (1990) Evaluation and management of scleroderma lung disease using bronchoalveolar lavage. *Am J Med* 88: 470–476.
26. Rubin LJ (1997) Primary pulmonary hypertension. *N Engl J Med* 336: 111–117.
27. Zhernakova A, Stahl EA, Trynka G, Raychaudhuri S, Festen EA, et al. (2011) Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. *PLoS Genet* 7: e1002004. doi:10.1371/journal.pgen.1002004.
28. Valdes AM, Thomson G (2009) Several loci in the HLA class III region are associated with T1D risk after adjusting for DRB1-DQB1. *Diabetes Obes Metab* 11(Suppl 1): 46–52.
29. Kochi Y, Yamada R, Kobayashi K, Takahashi A, Suzuki A, et al. (2004) Analysis of single-nucleotide polymorphisms in Japanese rheumatoid arthritis patients shows additional susceptibility markers besides the classic shared epitope susceptibility sequences. *Arthritis Rheum* 50: 63–71.
30. Tazi-Ahmini R, Cork MJ, Wengraf D, Wilson AG, Gawkrödger DJ, et al. (2003) Notch4, a non-HLA gene in the MHC is strongly associated with the most severe form of alopecia areata. *Hum Genet* 112: 400–403.
31. Petukhova L, Duvic M, Hordinsky M, Norris D, Price V, et al. (2010) Genome-wide association study in alopecia areata implicates both innate and adaptive immunity. *Nature* 466: 113–117.
32. Assassi S, Mayes MD, Arnett FC, Gourh P, Agarwal SK, et al. (2010) Systemic sclerosis and lupus: points in an interferon-mediated continuum. *Arthritis Rheum* 62: 589–598.
33. Trinchieri G (2010) Type I interferon: friend or foe? *J Exp Med* 207(10): 2053–2063.
34. Eloranta ML, Franck-Larsson K, Lovgren T, Kalamajski S, Ronnblom A, et al. (2010) Type I interferon system activation and association with disease manifestations in systemic sclerosis. *Ann Rheum Dis* 69: 1396–1402.
35. Zhao J, Kong HJ, Li H, Huang B, Yang M, et al. (2006) IRF-8/interferon (IFN) consensus sequence-binding protein is involved in Toll-like receptor (TLR) signaling and contributes to the cross-talk between TLR and IFN- γ signaling pathways. *J Biol Chem* 281: 10073–10080.
36. Wang H, Lee CH, Qi C, Taylor P, Feng J, et al. (2008) IRF8 regulates B-cell lineage specification, commitment, and differentiation. *Blood* 112: 4028–4038.
37. De Jager PL, Jia X, Wang J, de Bakker PI, Ottoboni L, et al. (2009) Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nat Genet* 41: 776–782.
38. Lefebvre V, Behringer RR, de Crombrughe B (2001) L-Sox5, Sox6 and Sox9 control essential steps of the chondrocyte differentiation pathway. *Osteoarthritis Cartilage* 9(Suppl A): S69–75.
39. Bobick BE, Matsche AI, Chen FH, Tuan RS (2010) The ERK5 and ERK1/2 signaling pathways play opposing regulatory roles during chondrogenesis of adult human bone marrow-derived multipotent progenitor cells. *J Cell Physiol* 224: 178–186.
40. Osaki M, Tan L, Choy BK, Yoshida Y, Cheah KS, et al. (2003) The TATA-containing core promoter of the type II collagen gene (COL2A1) is the target of interferon- γ -mediated inhibition in human chondrocytes: requirement for Stat1 α , Jak1 and Jak2. *Biochem J* 369: 103–115.
41. Kanno Y, Levi BZ, Tamura T, Ozato K (2005) Immune cell-specific amplification of interferon signaling by the IRF-4/8-PU.1 complex. *J Interferon Cytokine Res* 25: 770–779.
42. Nantel A, Mohammad-Ali K, Sherk J, Posner BI, Thomas DY (1998) Interaction of the Grb10 adapter protein with the Raf1 and MEK1 kinases. *J Biol Chem* 273: 10475–10484.
43. Simeon CP, Fonollosa V, Tolosa C, Palou E, Selva A, et al. (2009) Association of HLA class II genes with systemic sclerosis in Spanish patients. *J Rheumatol* 36: 2733–2736.
44. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, et al. (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet* 39: 870–874.
45. Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, et al. (2007) Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet* 39: 645–649.
46. (1980) Preliminary criteria for the classification of systemic sclerosis (scleroderma). Subcommittee for scleroderma criteria of the American Rheumatism Association Diagnostic and Therapeutic Criteria Committee. *Arthritis Rheum* 23: 581–590.
47. LeRoy EC, Black C, Fleischmajer R, Jablonska S, Krieg T, et al. (1988) Scleroderma (systemic sclerosis): classification, subsets and pathogenesis. *J Rheumatol* 15: 202–205.
48. Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 39: 906–913.
49. Skol AD, Scott LJ, Abecasis GR, Boehnke M (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* 38: 209–213.
50. Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21: 263–265.
51. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81: 559–575.

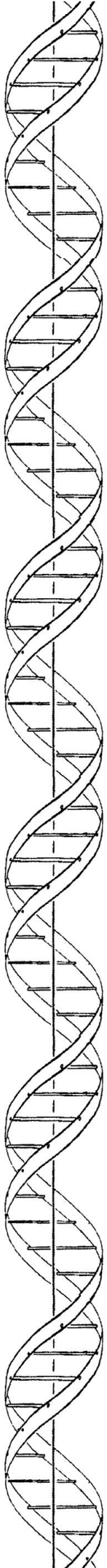


Table S1.

Chr.	Gene	SNP	Base Pair	Location	Change	Stage	N (case/control)	MAF (case/control)	P value†	OR (CI 95%)	Full set P‡	deSSc P†	ACCA+ P‡
12p12.1	<i>SOX5</i>	rs11047102	23,837,413	Intronic	T/C	GWAS	1400/5172	0.132/0.097	1.49x10 ⁻⁷	1.43 (1.26-1.63)	1.36x10 ⁻⁶	0.222	1.03x10 ⁻⁵
						Replication	1960/4971	0.108/0.102	0.244	1.08 (0.95-1.23)	0.162	0.351	0.00291
						Combined	3360/10143	0.118/0.099	5.11x10 ⁻⁶	1.24 (1.13-1.35)	7.52x10 ⁻⁶	0.127	1.39x10 ⁻⁷
16q24.1	<i>RPF8</i>	rs11642873	84,549,206	Intergenic	C/A	GWAS	1400/5172	0.144/0.197	1.39x10 ⁻⁷	0.72 (0.64-0.81)	3.89x10 ⁻⁶	0.151	0.000305
						Replication	1960/4971	0.143/0.186	6.88x10 ⁻⁷	0.78 (0.70-0.87)	5.81x10 ⁻⁷	0.00383	0.00176
						Combined	3360/10143	0.144/0.192	2.32x10 ⁻¹²	0.75 (0.69-0.81)	4.27x10 ⁻¹²	0.00305	1.38x10 ⁻⁶
14q21.1	---	rs12887070	42,027,287	Intergenic	A/C	GWAS	1400/5172	0.070/0.048	2.84x10 ⁻⁶	1.52 (1.28-1.81)	2.63x10 ⁻⁶	0.0438	0.00427
						Replication	1960/4971	0.055/0.055	8.79	1.01 (0.85-1.20)	0.126	0.00221	0.670
						Combined	3360/10143	0.061/0.051	0.000801	1.23 (1.09-1.39)	8.86x10 ⁻⁶	0.000251	0.0249
13q12.3	<i>UBL3</i>	rs7994117	29,386,799	Downstream	G/T	GWAS	1400/5172	0.227/0.183	3.15x10 ⁻⁶	1.28 (1.16-1.42)	3.34x10 ⁻⁵	0.157	8.39x10 ⁻⁶
						Replication	1960/4971	0.202/0.205	0.520	0.97 (0.88-1.07)	0.510	0.000893	0.589
						Combined	3360/10143	0.212/0.193	0.00490	1.11 (1.03-1.19)	0.0131	0.0474	0.000578
4p16.3	<i>DGKQ</i>	rs11724804	955,779	Intronic	A/G	GWAS	1400/5172	0.485/0.436	6.83x10 ⁻⁶	1.22 (1.12-1.33)	6.11x10 ⁻⁵	0.137	1.21x10 ⁻⁵
						Replication	1960/4971	0.467/0.451	0.137	1.06 (0.98-1.15)	0.00864	0.0147	0.703
						Combined	3360/10143	0.575/0.443	1.99x10 ⁻⁵	1.13 (1.07-1.20)	1.79x10 ⁻⁶	0.00477	0.000960
7p12.1	<i>GRB10</i>	rs12540874	50,632,416	Intronic	G/A	GWAS	1400/5172	0.462/0.409	3.00x10 ⁻⁶	1.23 (1.13-1.34)	0.000534	0.748	0.00169
						Replication	1960/4971	0.416/0.395	0.307	1.09 (1.01-1.18)	0.0183	0.395	0.0935
						Combined	3360/10143	0.435/0.403	1.27x10 ⁻⁶	1.15 (1.09-1.22)	2.76x10 ⁻⁵	0.397	0.000647
2q37.1	<i>AMRC9/PSMD1</i>	rs1868929	231,891,835	Intronic	T/C	GWAS	1400/5172	0.066/0.044	1.59x10 ⁻⁶	1.55 (1.30-1.86)	3.20x10 ⁻⁵	0.387	0.0139
						Replication	1960/4971	0.059/0.054	0.336	1.09 (0.92-1.28)	0.0164	0.000651	0.974
						Combined	3360/10143	0.062/0.048	7.67x10 ⁻⁵	1.28 (1.13-1.44)	3.27x10 ⁻⁶	0.00137	0.112

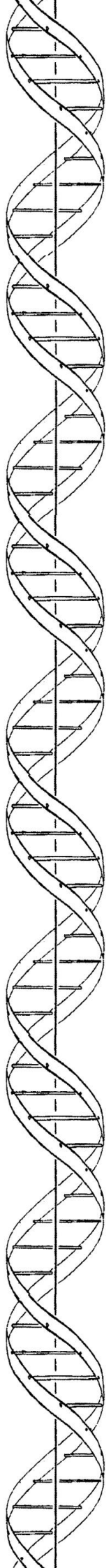


Table S2.

Chr.	Gene	SNP	Base Pair	Location	Change	Stage	N (case/control)	MAF (case/control)	P value†	OR (CI 95%)	Full set P‡	lcSSc P‡	ATA+ P‡
7p12.1	<i>AC009415.1</i>	rs2113648	52,029,517	Intergenic	A/G	GWAS	740/5172	0.277/0.225	6.21×10^{-6}	1.33 (1.18-1.51)	0.0128	0.958	0.00325
						Replication	959/4971	0.229/0.246	0.164	0.92 (0.81-1.04)	0.0660	0.109	0.661
12q13.2	<i>RPL41/ESYT1*</i>	rs11171747	54,804,675	Upstream	G/T	Combined	1699/10143	0.251/0.234	0.0363	1.10 (1.01-1.20)	0.651	0.266	0.0913
						GWAS	740/5172	0.446/0.385	2.19×10^{-6}	1.31 (1.17-1.46)	0.00176	0.433	0.00354
13q33.2	<i>EFNB2</i>	rs1477924	105,709,444	Intergenic	G/A	Replication	959/4971	0.408/0.372	0.00349	1.16 (1.05-1.29)	0.000462	0.00379	0.0189
						Combined	1699/10143	0.425/0.379	5.99×10^{-8}	1.23 (1.14-1.33)	1.76×10^{-6}	0.00792	0.000174
7q22.1	<i>ZNF789</i>	rs10235235	98,913,767	Intronic	C/T	GWAS	740/5172	0.230/0.180	9.16×10^{-6}	1.35 (1.19-1.54)	0.000801	0.132	0.000800
						Replication	959/4971	0.184/0.173	0.341	1.07 (0.93-1.22)	0.312	0.395	0.269
7p15.1	<i>JAZF1</i>	rs10275834	28,117,000	Intronic	T/C	Combined	1699/10143	0.205/0.177	0.000147	1.20 (1.09-1.31)	0.00158	0.0911	0.00166
						GWAS	740/5172	0.122/0.083	5.77×10^{-6}	1.50 (1.26-1.78)	0.0106	0.484	0.000168
7p15.1	<i>JAZF1</i>	rs10275834	28,117,000	Intronic	T/C	Replication	959/4971	0.090/0.091	0.802	0.98 (0.82-1.17)	0.111	0.0980	0.450
						Combined	1699/10143	0.104/0.087	0.00337	1.20 (1.06-1.36)	0.490	0.459	0.0457
7p15.1	<i>JAZF1</i>	rs10275834	28,117,000	Intronic	T/C	GWAS	740/5172	0.314/0.263	5.28×10^{-5}	1.28 (1.14-1.44)	2.35×10^{-5}	0.00803	0.0782
						Replication	959/4971	0.270/0.279	0.551	0.97 (0.86-1.08)	0.0520	0.0144	0.585
						Combined	1699/10143	0.289/0.270	0.0193	1.10 (1.02-1.18)	9.44×10^{-6}	0.000244	0.108

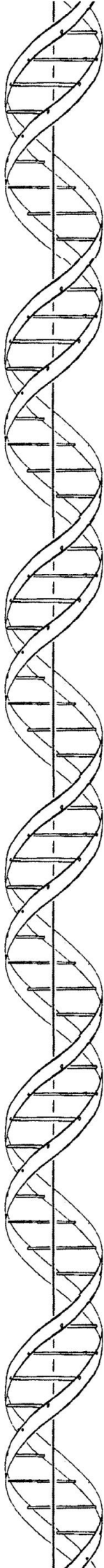


Table S3.

Chr.	Gene	SNP	Base Pair	Location	Change	Stage	N (case/control)	MAF (case/control)	P value†	OR (CI 95%)	Full set P‡	ATA+ P‡	lSSc P‡
1p13.3	<i>IL12RB2</i>	rs3790567	67,594,965	Intronic	A/G	GWAS	761/5172	0.307/0.250	3.64x10 ⁻⁶	1.34 (1.18-1.51)	2.97x10 ⁻⁵	0.0331	1.61x10 ⁻⁵
						Replication	1030/4971	0.272/0.255	0.458	1.04 (0.93-1.17)	0.00363	0.0663	0.0273
21q22.12	<i>RUNX1</i>	rs16993158	35,661,522	Intronic	C/T	GWAS	1791/10143	0.287/0.252	0.000197	1.17 (1.08-1.27)	3.39x10 ⁻⁷	0.00468	3.51x10 ⁻⁵
						Replication	761/5172	0.107/0.080	6.48x10 ⁻⁶	1.53 (1.27-1.83)	0.000334	0.330	0.000249
12p12.1	<i>SOX5</i>	rs11047102	23,837,413	Intronic	T/C	Combined	1791/10143	0.070/0.079	0.0916	0.85 (0.70-1.03)	0.344	0.0745	0.524
						GWAS	761/5172	0.086/0.077	0.0604	1.13 (0.99-1.29)	0.00115	0.0444	0.00250
						Replication	1030/4971	0.123/0.102	1.03x10 ⁻⁵	1.46 (1.24-1.72)	1.36x10 ⁻⁶	0.482	1.49x10 ⁻⁷
						Combined	1791/10143	0.127/0.099	0.00291	1.27 (1.09-1.48)	0.162	0.154	0.244
									1.39x10 ⁻⁷	1.36 (1.21-1.52)	7.52x10 ⁻⁶	0.121	5.11x10 ⁻⁶

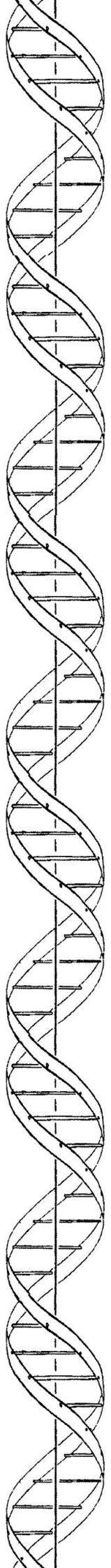


Table S4.

Chr.	Gene	SNP	Base Pair	Location	Change	Stage	N (case/control)	MAF (case/control)	P value†	OR (CI 95%)	Full set P‡	ACA+ P‡	dcSSc P‡
14q21.1	<i>LRFN5</i>	rs1959429	40,973,264	Intergenic	T/C	GWAS	447/5172	0.172/0.245	8.97x10 ⁻⁷	0.63 (0.53-0.75)	0.0189	0.106	0.0883
						Replication	626/4971	0.277/0.249	0.0918	1.13 (0.98-1.30)	0.00322	0.0353	0.0918
15q14	<i>ATPBD4/ZNF770</i>	rs4924647	33,282,199	Intergenic	T/G	GWAS	1073/10143	0.232/0.245	0.0371	0.89 (0.80-0.99)	0.634	0.647	0.0371
						Replication	447/5172	0.086/0.050	2.76x10 ⁻⁶	1.85 (1.44-2.38)	0.0130	0.746	0.00138
4q35.1	<i>DCTD/ODZ3</i>	rs4861533	184,028,292	Intergenic	C/A	GWAS	1073/10143	0.062/0.061	0.938	1.01 (0.78-1.31)	0.364	0.407	0.938
						Replication	626/4971	0.073/0.055	0.00172	1.33 (1.11-1.59)	0.288	0.670	0.00172
12p12.1	<i>SOX5</i>	rs9634098	24,097,345	Intergenic	T/C	GWAS	447/5172	0.301/0.231	1.50x10 ⁻⁶	1.46 (1.26-1.70)	0.0319	0.535	0.00686
						Replication	626/4971	0.243/0.238	0.746	1.02 (0.89-1.18)	0.483	0.460	0.746
12p12.1	<i>SOX5</i>	rs9634098	24,097,345	Intergenic	T/C	GWAS	1073/10143	0.267/0.234	0.000424	1.20 (1.09-1.34)	0.0412	0.884	0.000424
						Replication	447/5172	0.077/0.041	3.33x10 ⁻⁶	1.92 (1.46-2.51)	0.00689	0.0133	0.0367
						Combined	626/4971	0.043/0.045	0.247	0.84 (0.62-1.13)	0.298	0.539	0.247
						Combined	1073/10143	0.057/0.043	0.0265	1.25 (1.02-1.53)	0.238	0.233	0.0265

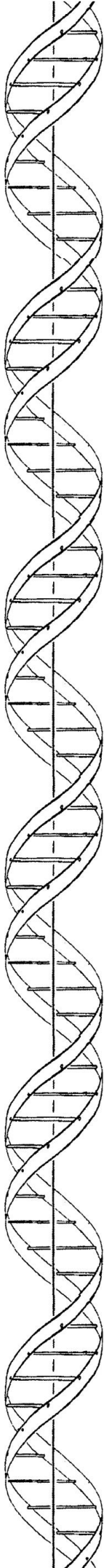


Table S5.

Phenotype	N		λ					OR 1.50							OR 1.30						
	Cases	Controls	All	US	Spain	Germany	Netherlands	MAF	MAF	MAF	MAF	MAF	MAF	MAF	MAF	MAF	MAF	MAF	MAF	MAF	
SSc	5,471	10,143	1.081	1.090	1.033	1.126	1.045	100	100	100	100	100	100	100	100	100	100	100	100	100	94
lcSSc	3,360	10,143	1.058	1.071	1.033	1.101	1.028	100	100	100	100	100	100	100	100	100	100	100	100	100	64
dcSSc	1,699	10,143	1.034	1.049	1.032	1.061	1.020	100	100	100	100	100	93	29	94	88	64	39	12	14	12
ACA+	1,791	10,143	1.050	1.051	1.042	1.081	1.012	100	100	100	100	100	95	33	96	91	69	43	14	14	14
ATA+	1,073	10,143	1.061	1.035	1.025	1.078	1.030	100	100	98	89	53	5	62	47	22	9	9	2	2	2

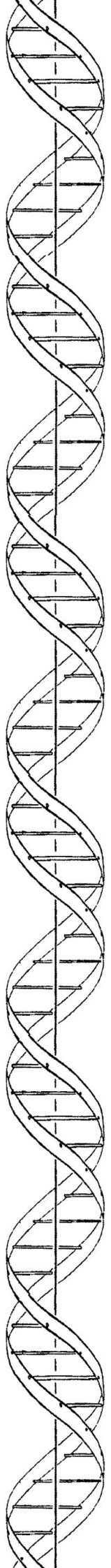


Table S6.

SNP	BP	<i>P</i> Value†	OR	Conditioned to rs9275390		Conditioned to rs6457617		Conditioned to rs443198	
				<i>P</i> value	OR	<i>P</i> value	OR	<i>P</i> value	OR
rs9275390	32,777,134	2.62×10^{-54}	2.385	NA	NA	1.52×10^{-67}	1.972	1.14×10^{-130}	2.267
rs6457617	32,771,829	1.99×10^{-36}	0.477	5.14×10^{-67}	0.703	NA	NA	6.83×10^{-105}	0.513
rs443198	32,298,384	8.84×10^{-21}	0.556	8.11×10^{-12}	0.691	7.43×10^{-11}	0.662	NA	NA

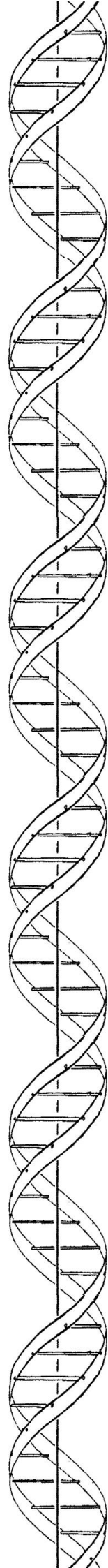


Table S7.

SNP	BP	P Value†	OR	Conditioned to rs129882		Conditioned to rs987870		Conditioned to rs135021		Conditioned to rs129763		Conditioned to rs9296015		Conditioned to rs1810472		Conditioned to rs6901221	
				P value	OR	P value	OR	P value	OR	P value	OR	P value	OR	P value	OR	P value	OR
rs129882	32,517,508	1.89x10 ⁻²⁷	2.168	NA	NA	3.89x10 ⁻²⁶	2.359	2.95x10 ⁻²⁷	2.097	2.39x10 ⁻²³	2.014	2.23x10 ⁻²⁹	2.038	1.19x10 ⁻³⁶	2.170	1.69x10 ⁻³⁰	2.166
rs987870	33,150,858	2.41x10 ⁻²⁰	2.093	6.01x10 ⁻¹⁶	2.564	NA	NA	7.43x10 ⁻²⁵	2.501	1.09x10 ⁻¹⁵	2.144	3.55x10 ⁻¹⁵	2.295	7.73x10 ⁻¹⁰	1.988	8.51x10 ⁻¹⁹	2.274
rs135021	33,153,536	1.95x10 ⁻¹²	1.656	1.47x10 ⁻⁶⁰	1.658	3.18x10 ⁻¹²¹	1.935	NA	2.50x10 ⁻⁵⁷	1.657	2.33x10 ⁻⁷⁷	1.745	6.07x10 ⁻⁵⁵	1.568	2.78x10 ⁻⁸⁸	1.783	
rs129763	32,698,903	1.47x10 ⁻¹¹	1.647	9.42x10 ⁻¹⁴⁹	1.442	3.04x10 ⁻¹⁷⁶	1.596	4.47x10 ⁻¹⁷⁹	1.604	NA	NA	0	1.574	9.19x10 ⁻¹⁸⁷	1.723	7.61x10 ⁻¹¹²	1.726
rs9296015	32,326,967	1.14x10 ⁻⁸	0.545	2.07x10 ⁻⁷	0.706	1.17x10 ⁻¹⁴	0.546	1.49x10 ⁻¹⁸	0.561	9.98x10 ⁻¹¹	0.633	NA	NA	1.27x10 ⁻¹⁶	0.566	2.63x10 ⁻¹³	0.594
rs1810472	33,191,099	2.14x10 ⁻⁸	1.498	1.38x10 ⁻³⁶	1.583	2.21x10 ⁻⁸	1.251	7.33x10 ⁻²⁶	1.363	8.48x10 ⁻⁴⁵	1.604	3.31x10 ⁻³⁶	1.570	NA	NA	4.96x10 ⁻³⁹	1.513
rs6901221	33,206,254	2.54x10 ⁻⁸	1.606	1.51x10 ⁻⁹	1.612	9.73x10 ⁻²³	1.613	2.23x10 ⁻¹³	1.659	7.49x10 ⁻¹⁶	1.643	1.23x10 ⁻⁹	1.516	9.47x10 ⁻¹²	1.504	NA	NA

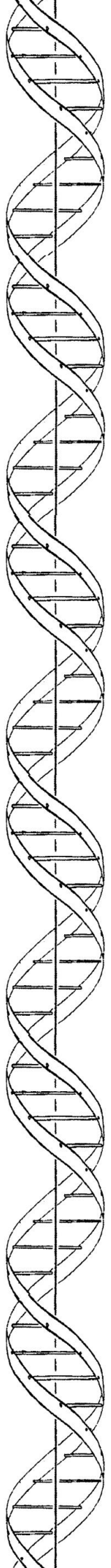


Table S8

SNP	Gene	Location	Change	Population	MAF (case/control)	<i>P</i> †	OR (CI 95%)
rs443198	<i>NOTCH4</i>	Exon	C/T	Spain	0.295/0.414	0.000190	0.593 (0.45-0.78)
				Germany	0.240/0.339	0.00236	0.615 (0.45-0.84)
				Netherlands	0.233/0.327	0.0692	0.624 (0.37-1.04)
				US	0.243/0.380	3.40x10 ⁻¹⁵	0.524 (0.45-0.62)
rs6457617	<i>HLA-DQB1</i>	Intergenic	C/T	Spain	0.325/0.474	4.81x10 ⁻⁶	0.535 (0.41-0.70)
				Germany	0.298/0.504	3.40x10 ⁻⁹	0.417 (0.31-0.56)
				Netherlands	0.430/0.511	0.147	0.723 (0.47-1.12)
				US	0.302/0.488	9.38x10 ⁻²⁵	0.455 (0.39-0.53)
rs9275390	<i>HLA-DQB1</i>	Intergenic	C/T	Spain	0.458/0.306	1.20x10 ⁻⁶	1.920 (1.47-2.50)
				Germany	0.475/0.250	6.45x10 ⁻¹³	2.733 (2.06-3.62)
				Netherlands	0.361/0.269	0.0664	1.530 (0.97-2.42)
				US	0.456/0.245	1.09x10 ⁻³⁹	2.585 (2.24-2.99)

Table S9.

SNP	Gene	Location	Change	Population	MAF (case/control)	<i>P</i> †	OR (CI 95%)
rs9296015	<i>NOTCH4</i>	Intergenic	A/G	Spain	0.161/0.280	0.001614	0.492 (0.31-0.77)
				Germany	0.077/0.174	0.00145	0.399 (0.22-0.72)
				Netherlands	0.148/0.187	0.3581	0.754 (0.42-1.38)
				US	0.111/0.178	0.000217	0.580 (0.43-0.78)
rs3129882	<i>HLA-DRA</i>	Intron	G/A	Spain	0.593/0.448	0.000804	1.793 (1.27-2.53)
				Germany	0.691/0.444	1.64x10 ⁻⁹	2.792 (1.98-3.94)
				Netherlands	0.580/0.388	0.000393	2.173 (1.40-3.37)
				US	0.632/0.448	6.35x10 ⁻¹⁵	2.117 (1.75-2.57)
rs3129763	<i>HLA-DQA1/DRB1</i>	Intergenic	A/G	Spain	0.340/0.231	0.00356	1.716 (1.19-2.48)
				Germany	0.363/0.245	0.000952	1.759 (1.25-2.47)
				Netherlands	0.330/0.239	0.0550	1.568 (0.99-2.49)
				US	0.349/0.250	1.81x10 ⁻⁶	1.606 (1.32-1.96)
rs987870	<i>HLA-DPA1/DPB1</i>	Intron	C/T	Spain	0.333/0.191	6.66x10 ⁻⁵	2.112 (1.46-3.08)
				Germany	0.310/0.133	1.92x10 ⁻⁹	2.926 (2.04-4.21)
				Netherlands	0.250/0.156	0.0206	1.804 (1.09-2.99)
				US	0.240/0.142	6.25x10 ⁻⁹	1.907 (1.53-2.38)
rs3135021	<i>HLA-DPA1/DPB1</i>	Intron	A/G	Spain	0.438/0.328	0.00722	1.601 (1.13-2.26)
				Germany	0.369/0.271	0.00769	1.574 (1.13-2.20)
				Netherlands	0.363/0.280	0.0934	1.468 (0.94-2.31)
				US	0.410/0.285	8.29x10 ⁻⁹	1.738 (1.44-2.10)
rs6901221	<i>HLA-DPA1/DPB1</i>	Intron	C/A	Spain	0.160/0.103	0.0352	1.667 (1.03-2.69)
				Germany	0.244/0.172	0.0212	1.558 (1.07-2.28)
				Netherlands	0.227/0.180	0.267	1.340 (0.80-2.25)
				US	0.235/0.156	4.85x10 ⁻⁶	1.668 (1.38-2.08)

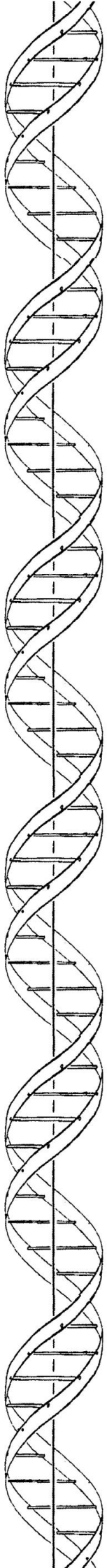


Table S10.

Chr.	Gene	SNP	Base Pair	Change	lcSSc		dcSSc		ACA+		ATA+		Refs
					<i>P</i> [†] value	OR (95% CI)	<i>P</i> [†] value	OR (95% CI)	<i>P</i> [†] value	OR (95% CI)	<i>P</i> [†] value	OR (95% CI)	
1	<i>CD247</i>	rs2056626	165,687,049	G/T	2.66x10 ⁻⁶	0.81 (0.75-0.89)	7.70x10 ⁻³	0.85 (0.78-0.96)	7.10x10 ⁻³	0.86 (0.77-0.96)	3.22x10 ⁻⁵	0.74 (0.64-0.85)	[12]
1	<i>TNFSF4</i>	rs2205960	171,458,098	T/G	7.70x10 ⁻⁴	1.18 (1.03-1.31)	0.506	1.05 (0.92-1.19)	0.0162	1.17 (1.03-1.33)	9.05x10 ⁻³	1.23 (1.05-1.45)	[13,14]
2	<i>STAT4</i>	rs3821236	191,611,003	A/G	8.86x10 ⁻⁸	1.31 (1.19-1.48)	5.96x10 ⁻⁴	1.25 (1.10-1.43)	1.18x10 ⁻⁴	1.29 (1.47-4.37)	1.53x10 ⁻³	1.30 (1.11-1.52)	[4, 6, 7, 8]
4	<i>BANK1</i>	rs10516487	102,970,099	T/C	0.317	1.05 (0.96-1.15)	0.0103	0.85 (0.75-0.96)	0.140	1.09 (0.97-1.23)	0.109	0.88 (0.75-1.03)	[5, 9]
7	<i>IRF5</i>	rs10488631	128,381,419	C/T	1.64x10 ⁻¹⁰	1.50 (1.32-1.69)	1.27x10 ⁻⁹	1.61 (1.38-1.88)	1.88x10 ⁻⁷	1.52 (1.30-1.79)	8.25x10 ⁻⁷	1.63 (1.34-1.98)	[15]
8	<i>BLK</i>	rs2736340	11,381,382	T/C	1.54x10 ⁻⁴	1.20 (1.09-1.32)	1.39x10 ⁻³	1.22 (1.08-1.38)	1.45x10 ⁻⁴	1.27 (1.12-1.44)	0.387	1.07 (0.92-1.26)	[10, 11]

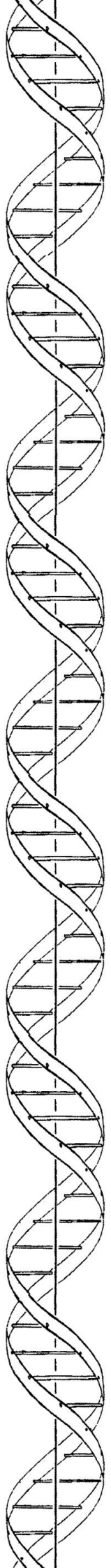


Table S11.

Population	Population Size		Sex (cases/controls)		Subtype		ACA Positive	ATA Positive
	Cases	Controls	Female	Male	Diffuse	Limited		
Overall	5471	10143	0.86/0.79	0.14/0.21	0.34	0.66	0.37	0.22
GWAS SSc cohorts								
Spain	364	384	0.88/0.75	0.10/0.25	0.30	0.64	0.42	0.20
Germany	270	671	0.88/0.62	0.11/0.38	0.40	0.55	0.42	0.31
The Netherlands	176	639	0.72/0.51	0.28/0.49	0.28	0.51	0.22	0.26
US	1486	3478	0.88/0.88	0.12/0.12	0.36	0.64	0.32	0.17
Replication SSc cohorts								
US	616	1143	0.88/0.44	0.12/0.56	0.41	0.59	0.32	0.15
Belgium	187	272	0.78/0.45	0.22/0.54	0.33	0.67	0.33	0.25
Italy	500	509	0.92/0.65	0.08/0.35	0.26	0.74	0.42	0.33
Sweden	268	278	0.80/0.78	0.20/0.22	0.27	0.73	0.28	0.18
UK	485	380	0.84/0.44	0.16/0.56	0.27	0.73	0.38	0.15
Norway	113	282	0.85/---	0.15/---	0.35	0.65	0.54	0.14
Spain	626	705	0.87/0.60	0.11/0.37	0.32	0.68	0.47	0.23
The Netherlands	204	350	0.71/0.45	0.29/0.54	0.32	0.68	0.25	0.27
Germany	176	291	0.81/0.44	0.13/0.27	0.42	0.58	0.39	0.32

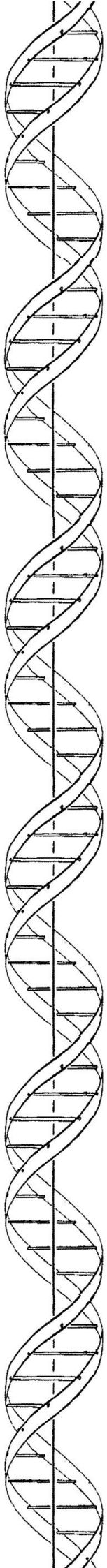


Figure S1.

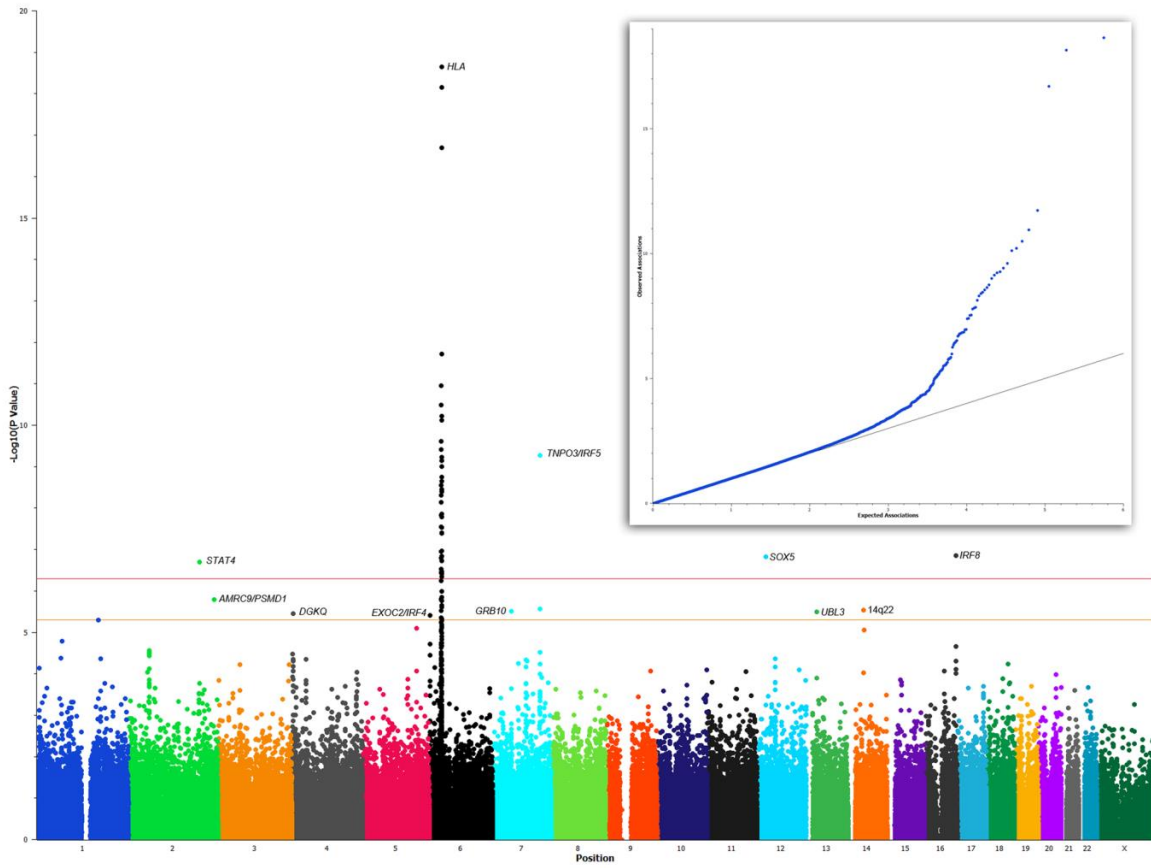


Figure S2.

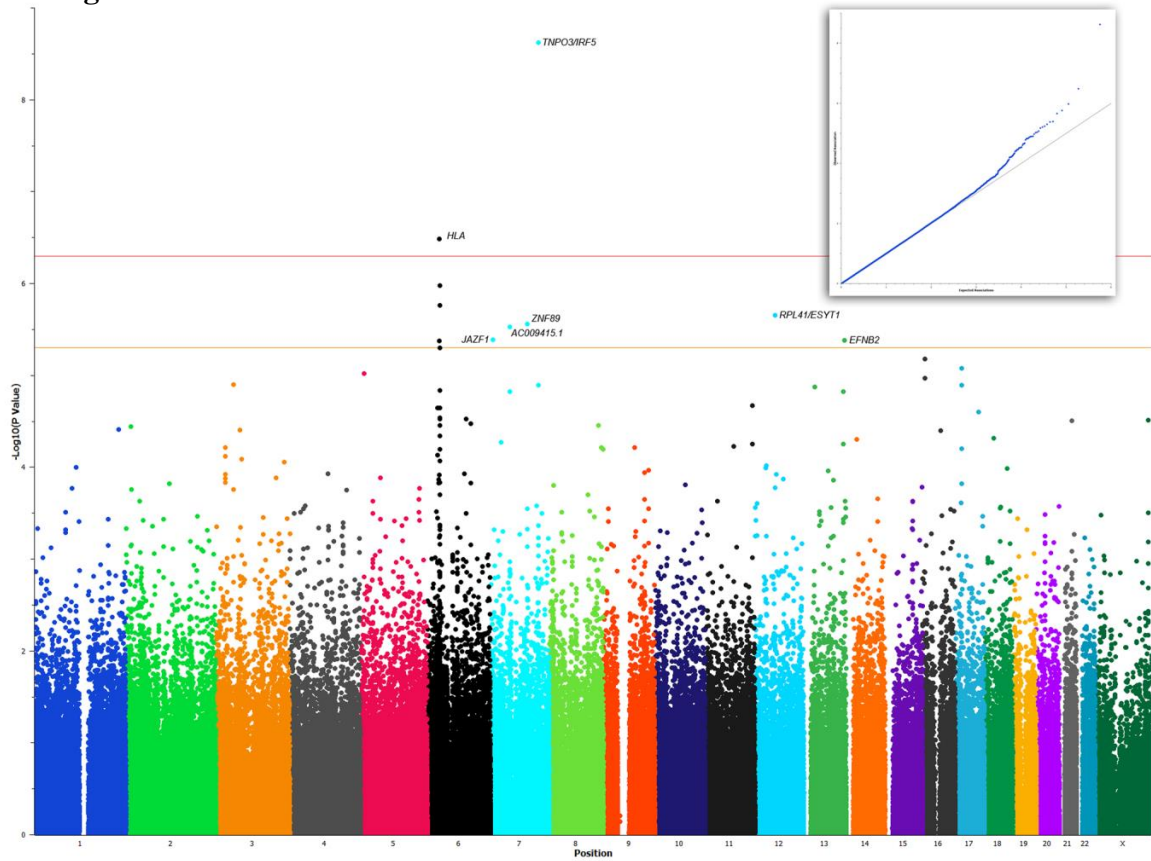


Figure S3.

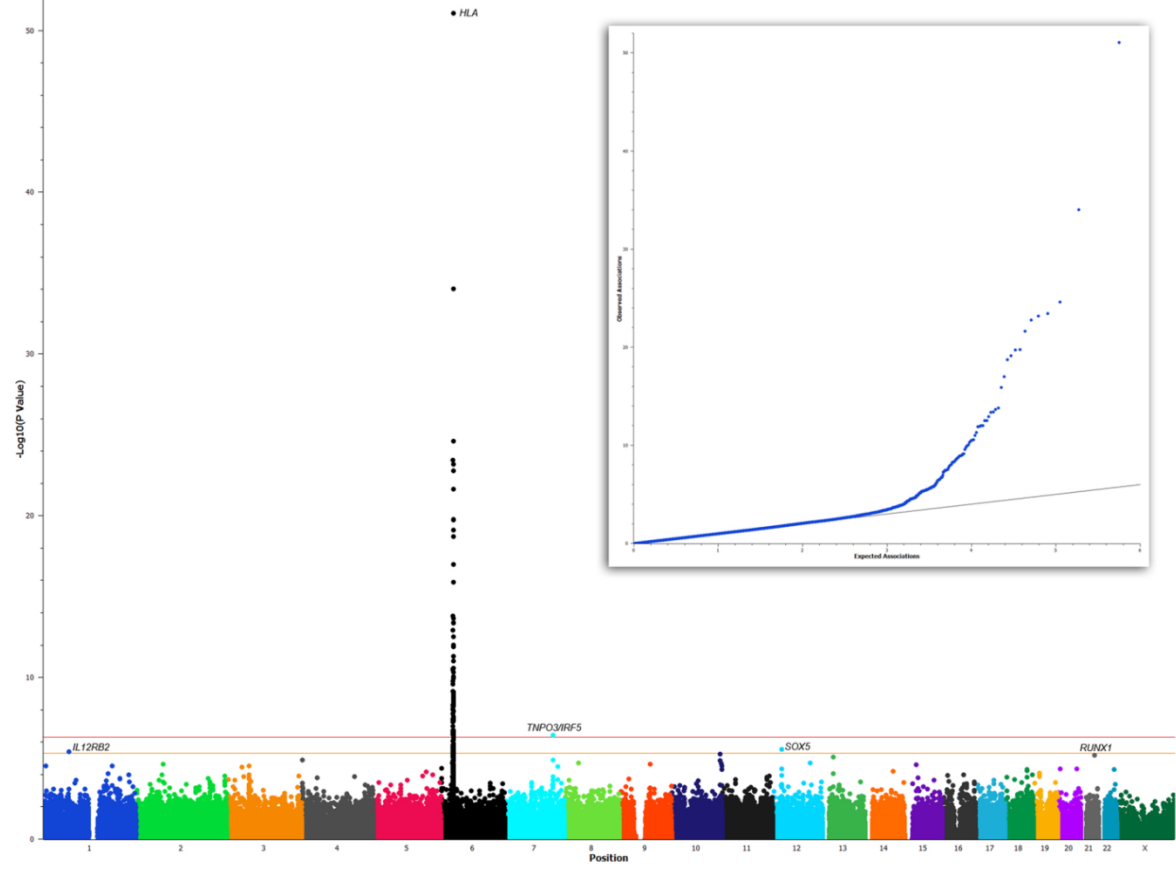


Figure S4.

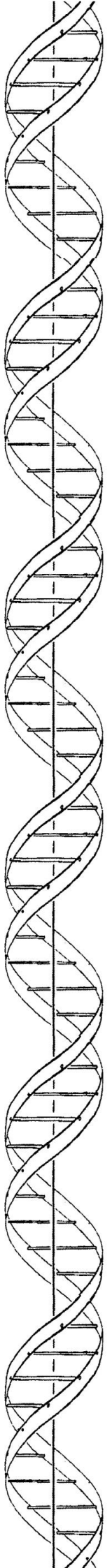
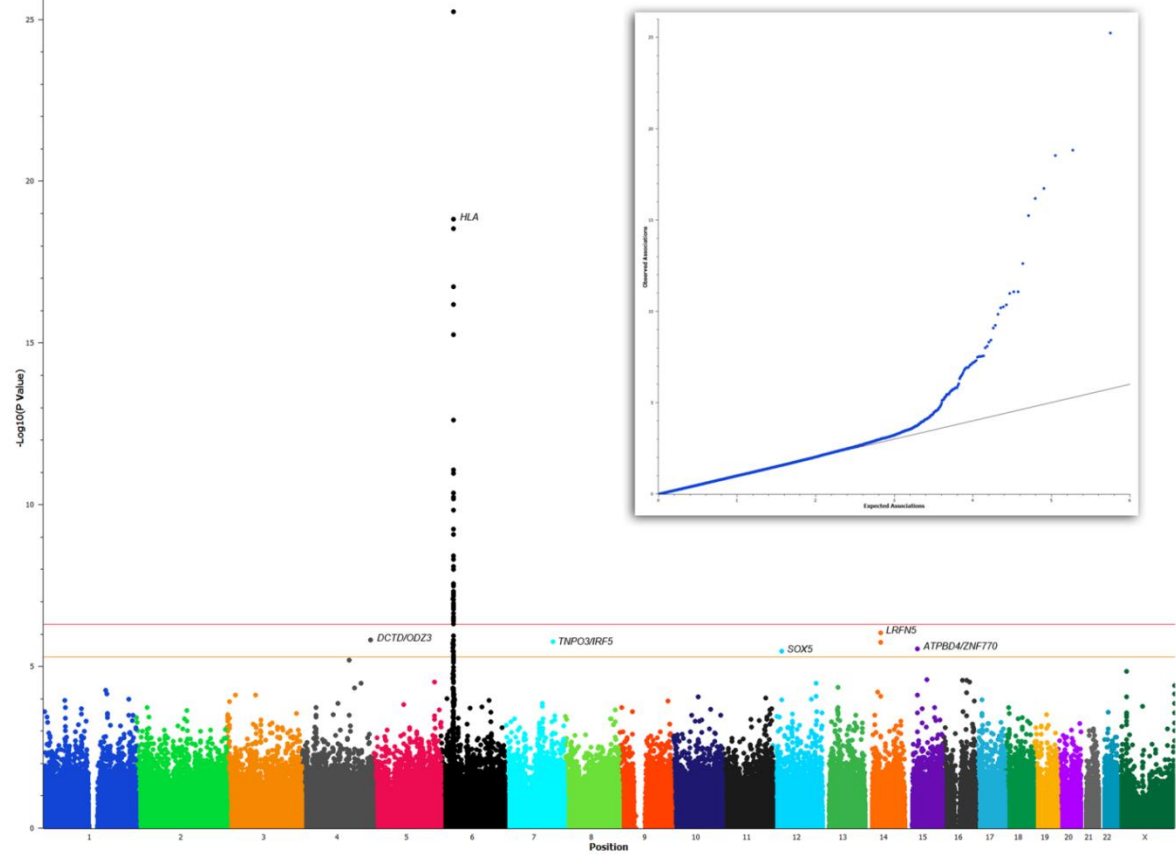
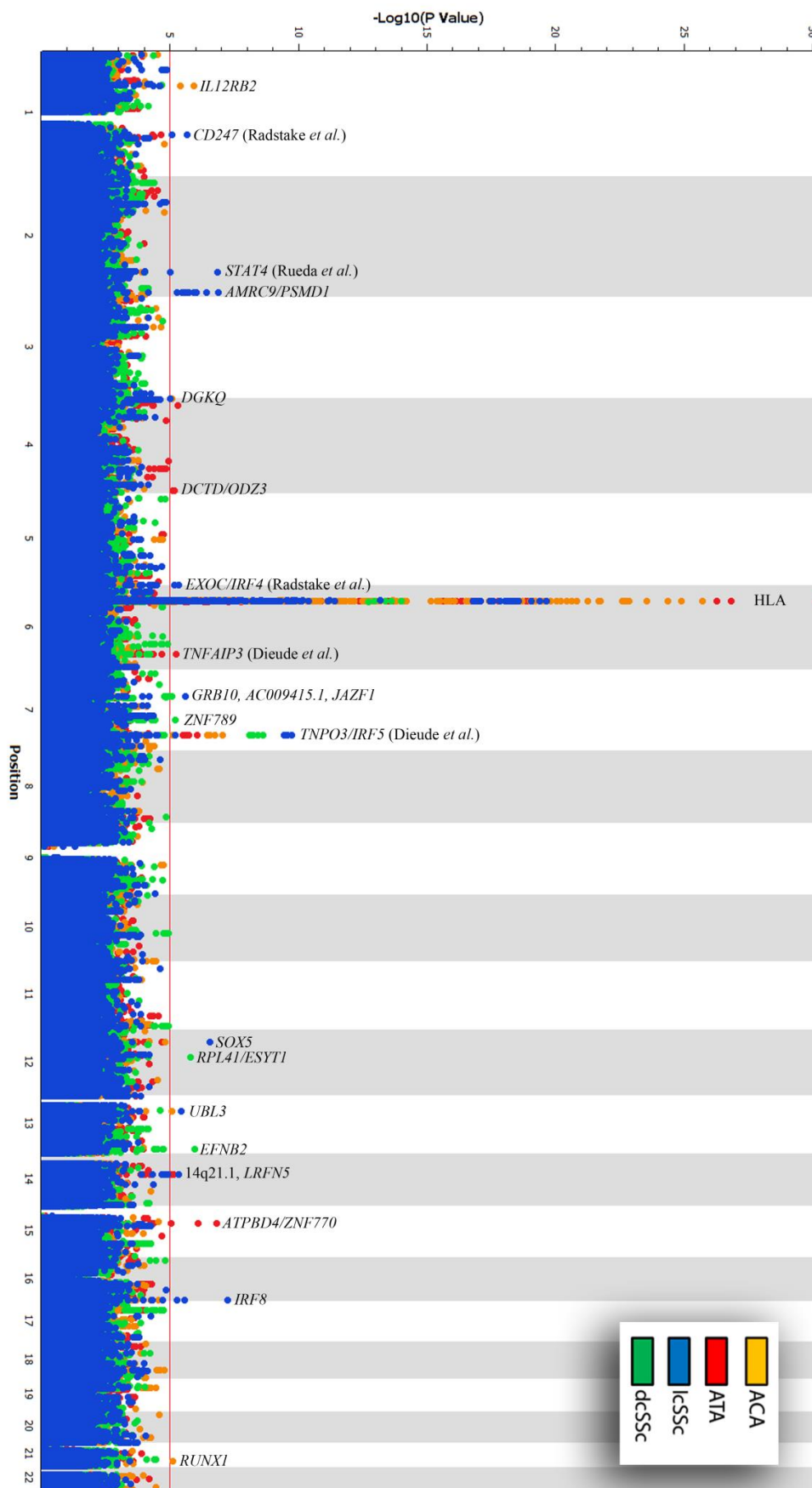
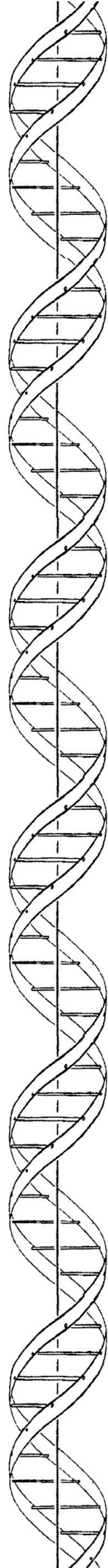
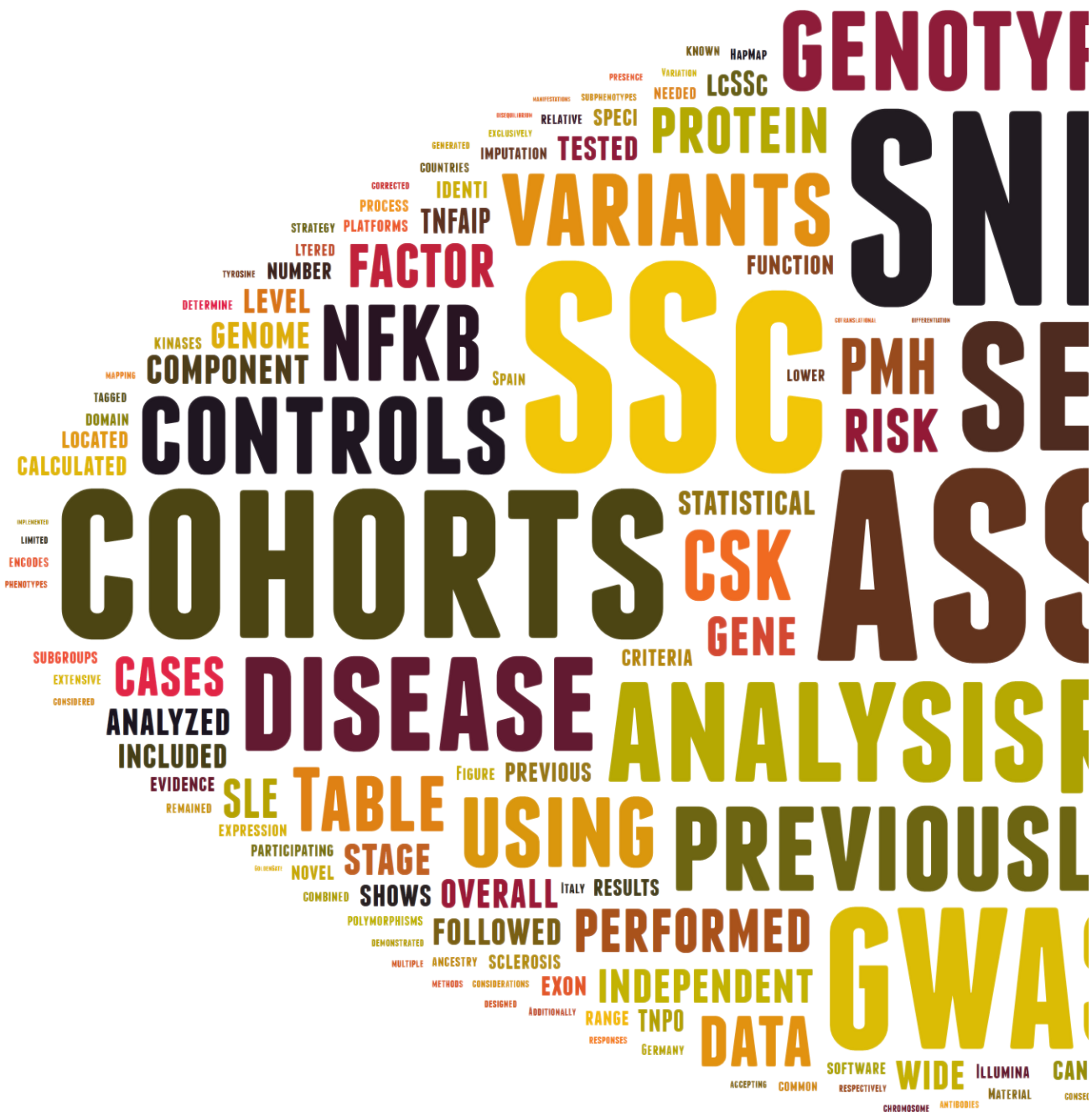
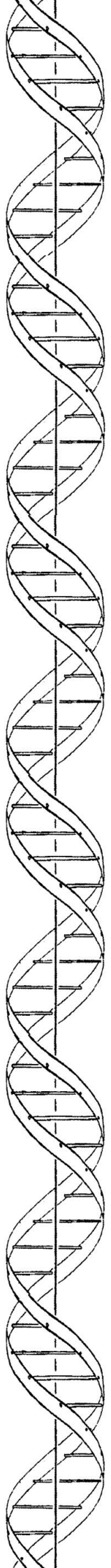
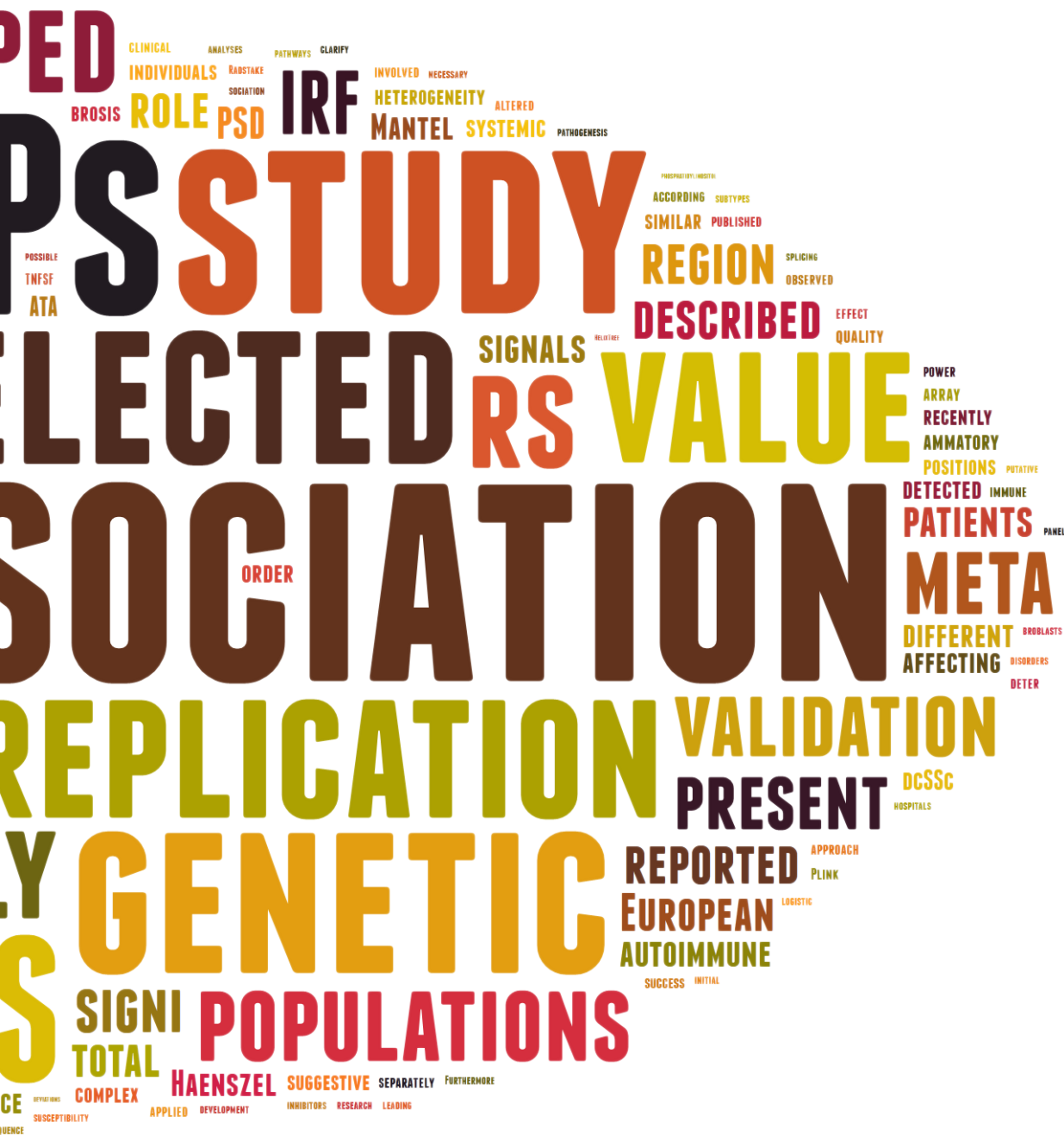


Figure S5.

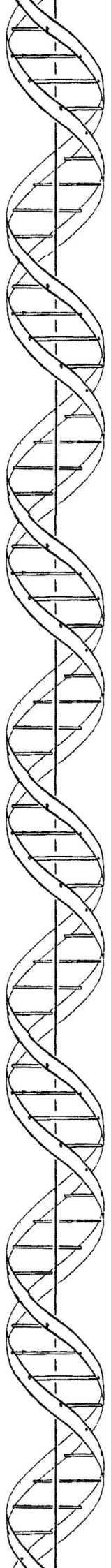








IDENTIFICATION OF *CSK* AS A SYSTEMIC SCLEROSIS
 GENETIC RISK FACTOR THROUGH GENOME WIDE
 ASSOCIATION STUDY FOLLOW-UP. *HUMAN
 MOLECULAR GENETICS*, 2012.



Identification of *CSK* as a systemic sclerosis genetic risk factor through Genome Wide Association Study follow-up

Jose-Ezequiel Martin^{1,*}, Jasper C. Broen², F. David Carmona¹, Maria Teruel¹, Carmen P. Simeon³, Madelon C. Vonk², Ruben van 't Slot⁴, Luis Rodriguez-Rodriguez⁵, Esther Vicente⁶, Vicente Fonollosa³, Norberto Ortego-Centeno⁷, Miguel A. González-Gay⁸, Francisco J. García-Hernández⁹, Paloma García de la Peña¹⁰, Patricia Carreira¹¹, Spanish Scleroderma Group[†], Alexandre E. Voskuyl¹², Annemie J. Schuerwegh¹³, Piet L.C.M. van Riel², Alexander Kreuter¹⁴, Torsten Witte¹⁵, Gabriella Riemekasten¹⁶, Paolo Airo¹⁷, Raffaella Scorza¹⁸, Claudio Lunardi¹⁹, Nicolas Hunzelmann²⁰, Jörg H.W. Distler²¹, Lorenzo Beretta¹⁸, Jacob van Laar²², Meng May Chee²³, Jane Worthington²⁴, Ariane Herrick²⁴, Christopher Denton²⁵, Filemon K. Tan²⁶, Frank C. Arnett²⁶, Shervin Assassi²⁶, Carmen Fonseca²⁵, Maureen D. Mayes²⁶, Timothy R.D.J. Radstake^{2,‡}, Bobby P.C. Koeleman^{4,‡} and Javier Martin^{1,‡}

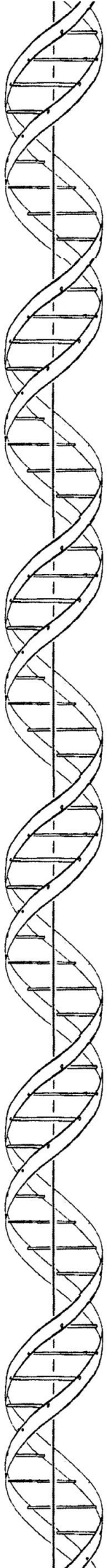
¹Instituto de Parasitología y Biomedicina Lopez-Neyra, CSIC, Granada, Spain, ²Department of Rheumatology, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands, ³Servicio de Medicina Interna, Hospital Valle de Hebron, Barcelona, Spain, ⁴Department of Medical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands, ⁵Servicio de Reumatología, Hospital Clínico San Carlos, Madrid, Spain, ⁶Servicio de Reumatología, Hospital La Princesa, Madrid, Spain, ⁷Servicio de Medicina Interna, Hospital Clínico Universitario, Granada, Spain, ⁸Servicio de Reumatología, Hospital Marqués de Valdecilla, Santander, Spain, ⁹Servicio de Medicina Interna, Hospital Virgen del Rocío, Sevilla, Spain, ¹⁰Servicio de Reumatología, Hospital Ramón y Cajal, Madrid, Spain, ¹¹Hospital 12 de Octubre, Madrid, Spain, ¹²VU University Medical Center, Amsterdam, The Netherlands, ¹³Department of Rheumatology, University of Leiden, Leiden, The Netherlands, ¹⁴Department of Dermatology, Josefs-Hospital, Ruhr University Bochum, Bochum, Germany, ¹⁵Hannover Medical School, Hannover, Germany, ¹⁶Department of Rheumatology and Clinical Immunology, Charité University Hospital, Berlin, Germany, ¹⁷Rheumatology Unit and Chair, Spedali Civili, Università de gli Studi, Brescia, Italy, ¹⁸Referral Center for Systemic Autoimmune Diseases, Fondazione IRCCS Ca'Granda Ospedale MaRepiore Policlinico and University of Milan, Milan, Italy, ¹⁹Department of Medicine, Policlinico GB Rossi, University of Verona, Verona, Italy, ²⁰Department of Dermatology, University of Cologne, Cologne, Germany, ²¹Department of Internal Medicine 3, Institute for Clinical Immunology, University of Erlangen-Nuremberg, Erlangen 91054, Germany, ²²Institute of Cellular Medicine, Newcastle University, Newcastle Upon Tyne, UK, ²³Centre for Rheumatic Diseases, Glasgow Royal Infirmary, Glasgow, UK, ²⁴Department of Rheumatology and Epidemiology, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK, ²⁵Centre for Rheumatology, Royal Free and University College School, London, UK and ²⁶The University of Texas Health Science Center–Houston, Houston, TX, USA

Received November 23, 2011; Revised and Accepted March 5, 2012

*To whom correspondence should be addressed. Email: cebercoto@ipb.csic.es

†See supplementary note.

‡These authors contributed equally to this work.



Systemic sclerosis (SSc) is complex autoimmune disease affecting the connective tissue; influenced by genetic and environmental components. Recently, we performed the first successful genome-wide association study (GWAS) of SSc. Here, we perform a large replication study to better dissect the genetic component of SSc. We selected 768 polymorphisms from the previous GWAS and genotyped them in seven replication cohorts from Europe. Overall significance was calculated for replicated significant SNPs by meta-analysis of the replication cohorts and replication-GWAS cohorts (3237 cases and 6097 controls). Six SNPs in regions not previously associated with SSc were selected for validation in another five independent cohorts, up to a total of 5270 SSc patients and 8326 controls. We found evidence for replication and overall genome-wide significance for one novel SSc genetic risk locus: *CSK* [P -value = 5.04×10^{-12} , odds ratio (OR) = 1.20]. Additionally, we found suggestive association in the loci *PSD3* (P -value = 3.18×10^{-7} , OR = 1.36) and *NFKB1* (P -value = 1.03×10^{-6} , OR = 1.14). Additionally, we strengthened the evidence for previously confirmed associations. This study significantly increases the number of known putative genetic risk factors for SSc, including the genes *CSK*, *PSD3* and *NFKB1*, and further confirms six previously described ones.

INTRODUCTION

Systemic sclerosis (scleroderma, SSc) is an autoimmune disease characterized by vascular damage, altered immune responses and extensive fibrosis of skin and internal organs (1), in which common genetic factors play an essential role, similar to most complex autoimmune diseases (2,3). So far, only a limited number of genes explaining little of the genetic variance present have been found in SSc (4,5).

In other autoimmune complex diseases for which extensive follow-up studies have been performed, such as inflammatory bowel disease and systemic lupus erythematosus (SLE), ~90 and 35 associated genes, respectively, have been identified (6,7). Therefore, it is expected that a number of risk factors for SSc are still to be defined. To date, only two genome-wide association studies (GWAS) have been published in populations of European ancestry in SSc (8,9). This calls for further replication studies and meta-analysis of SSc.

Due to the high rate of type 1 errors inherent to the GWAS technique, a number of strategies can be followed to discern truly associated genes from false positives in the tier 2 range of associations (ranging P -values from 10^{-3} to 5×10^{-8}), e.g. biological pathway analysis, meta-GWAS analysis or genetic interaction analysis. Another approach is to select the most strongly associated genetic variants from GWAS, where most true associations are harbored, by just accepting a small proportion of false negatives left out of the replication study. In this study, we performed the latter approach.

SSc is a clinically heterogeneous disease with a wide range of clinical manifestations (3); patients can be classified according to the severity of skin or organ involvement of the disease (10), or according to the presence or absence of several highly disease-specific auto-antibodies which are almost mutually exclusive in individual patients (1). Each of these disease phenotypes has proven to have specific genetic associations (11–17). In order to determine more of the genetic component of this profoundly disabling disease, such considerations need to be taken into account when selecting a battery of genetic variants to further test in other populations.

Considering the above, we followed the strategy of replication and validation of 768 genetic associations selected from our previous GWAS of SSc under different criteria in seven independent cohorts of European ancestry from five

European countries, and subsequently performed a meta-analysis including a total of 5270 SSc patients and 8326 healthy controls.

RESULTS

After quality filtering the replication stage data, 720 out of the 768 selected SNPs were analyzed. Mantel–Haenszel meta-analysis was performed for the three replication cohorts and for those cohorts together with the four GWAS cohorts. Six SNPs were selected for the validation stage in five independent cohorts. The overall process followed in the present study is summarized in Figure 1.

The genotyping success call rate was 99.88 and 97.99% in the replication and validation stages, respectively. Table 1 shows the statistics for the six SNPs selected for validation, while Table 2 shows previously described associations with SSc. Figure 2 shows meta-analysis association results of GWAS and replication cohorts for all 720 SNPs. Figure 3 shows relative risk for each of the three novel associations found in this study, for each population analyzed separately and for the meta-analysis. Detailed analysis and results for each associated region can be found below.

Novel SSc genetic associations

After the replication stage (genotyping of 768 SNPs selected from GWAS data), six SNPs were selected for validation with GWAS and replication combined Mantel–Haenszel P -value (P_{MH}) lower than 1×10^{-5} (in either the complete set of patients or its subphenotypes) and located in regions not previously associated with SSc. The SNPs in *NFKB1*, *PSD3*, *ACADS* and *CSK* were selected for their association with overall SSc, while those in *IPO5/FARPI* and *ADAMTS17* were selected for their association with the anti-topoisomerase autoantibody (ATA)-positive subgroup of patients. Out of the six SNPs, we were able to validate the association in one of them at the GWAS level ($P < 5 \times 10^{-8}$) and in another two at a suggestive level of association ($5 \times 10^{-8} < P < 5 \times 10^{-6}$).

The GWAS level association in the meta-analysis of all 11 analyzed cohorts was observed for single nucleotide

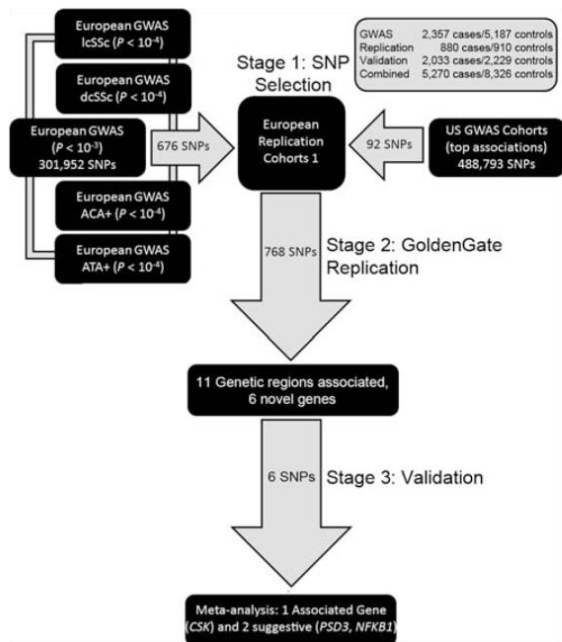


Figure 1. Schematic showing the overall process followed during the present study, along with the number of SNPs associated and considered in each step. lcSSc, limited cutaneous scleroderma; dcSSc, diffuse cutaneous scleroderma; ACA, anti-centromere autoantibody; ATA, anti-topoisomerase autoantibody.

polymorphisms (SNP) rs1378942 located in an intron of the *CSK* gene in chromosome 15 [P_{MH} value = 5.04×10^{-12} , odds ratio (OR) = 1.202 (1.14–1.27)] with SSc (Table 1). Furthermore, we observed a suggestive overall level of association in two of the selected SNPs in all GWAS, replication and validation cohorts. One of them was located in the *PSD3* gene, rs10096702 [P_{MH} value = 3.18×10^{-7} , OR = 1.363 (1.21–1.54)], while the other was in the *NFKB1* gene, rs1598859 [P_{MH} value = 1.03×10^{-6} , OR = 1.140 (1.08–1.20)] (Table 1). Interestingly, we were only able to validate associations with the overall disease, but none of the subphenotype specific ones.

Previously reported SSc genetic associations

Associations with SSc previously reported were found for *STAT4* (rs10168266, P_{MH} value = 1.81×10^{-7} , OR = 1.234), *CD247* (rs2056626, P_{MH} value = 1.14×10^{-8} , OR = 0.832), *TNFSF4* (rs4916334, P_{MH} value = 1.00×10^{-5} , OR = 0.861), *TNFAIP3* (rs2230926, P_{MH} value = 2.29×10^{-6} , OR = 1.463) and *TNPO3/IRF5* (rs4728142, P_{MH} value = 4.74×10^{-10} , OR = 1.216; and rs10488631, P_{MH} value = 1.58×10^{-18} , OR = 1.513), further confirming their role in SSc (Table 2). Since several SNPs within both *TNPO3/IRF5* and *IRF8* genomic regions were selected, a conditional logistic regression analysis was performed to determine the independent signals of both loci (Table 3). While these associations found within each previously reported region spanned across the whole disease, SNP rs11642873

in *IRF8* was significantly associated with lcSSc (P_{MH} value = 2.30×10^{-9} , OR = 0.730) and only marginally with dcSSc (P_{MH} value = 3.77×10^{-2} , OR = 0.874).

DISCUSSION

Through a large replication study, designed using previously published GWAS data, we have been able to identify novel genetic associations with SSc (Table 1), and to provide further evidence for previously reported associations (Table 2). In the present study, a genetic variant within the *CSK* region, selected for replication because it was associated with SSc in the US cohort, has been identified as susceptibility factor for this disease. Furthermore, two other suggestive associations have been found in *PSD3* and *NFKB1* genetic variants, which were included in this study because they reached statistical significance in the European cohort. Further investigation will be required to further clarify the potential role of these two signals in SSc genetic predisposition.

CSK (c-src tyrosine kinase) is known to phosphorylate a tyrosine at the C-terminus of src kinases leading to their inactivation (18,19). In turn, src kinases are involved in fibrosis through their regulation of *FAK* (19,20), which is necessary for transmission of integrin signaling upon adhesion of fibroblasts to the extracellular matrix [and thus, their differentiation into myfibroblasts (21)] and has been involved in experimental pulmonary fibrosis (20), a major hallmark of SSc. Indeed, it has been demonstrated that either incubation of fibroblasts with Csk inhibitors or overexpression of Csk lead to a decreased expression of *COL1A1*, *COL1A2* and *FNI*, which are key components of the fibrotic process (22). Thus, genetics variants in *CSK* can be affecting its expression or functionality in a way that src kinases are not inhibited, which in turn will contribute to the fibrosis in SSc. Furthermore, the *CSK* variant rs1378938, which is in relatively high linkage disequilibrium (LD) with the SSc-associated rs1378942 reported here ($r^2 = 0.72$ in the CEU population of the HapMap project), has been recently associated with celiac disease (23). Consequently, it is likely that *CSK* may represent another common autoimmunity risk factor. Further studies in related autoimmune disorders, such as SLE and rheumatoid arthritis (RA), may be performed to draw firm conclusions about this hypothesis.

NFKB has been extensively described to participate in and control the inflammatory process and thus, its role in the development of autoinflammatory disorders is widely accepted (24,25). The gene *NFKB1* (nuclear factor of kappa light polypeptide gene enhancer in B-cell 1) encodes a 105 kDa protein which can undergo cotranslational processing by the 26S proteasome to produce a 50 kDa protein. The 105 kDa protein is a Rel protein-specific transcription inhibitor and the 50 kDa protein is a DNA-binding subunit of the *NFKB* protein complex. SNP rs1598859, located in an intron of *NFKB1*, has been identified in this study as a risk genetic factor for SSc. This variant, or any other in the same haplotypic block, could be affecting the expression or the function of *NFKB1*, altering the inflammatory response, and thus participating in the development and course of the disease. Indeed, an ATTG in/del in the promoter of *NFKB1* has been recently

Table 1. Novel genetic associations found in this study with SSc

Phenotype	Chr.	Gene	SNP	Base pair ^a	Location	Change ^b	Stage	Sample size Cases	Sample size Controls	MAF	P-value ^c	OR ^d	CI (95%)	Breslow-Day P-value
SSc	4q24	<i>NFKB1</i>	rs1598859	103 725 482	Intron	C/T	GWAS	2357	5187	0.358	3.12×10^{-6}	1.188	1.11–1.28	0.151
							GWAS + Rep	3237	6097	0.354	6.51×10^{-6}	1.159	1.09–1.24	0.043
							GWAS + Rep + Val	5270	8326	0.356	1.03×10^{-6}	1.140	1.08–1.20	0.032
SSc	8p22	<i>PSD3</i>	rs10096702	18 642 877	Intron	A/G	GWAS	2357	5187	0.044	1.05×10^{-5}	1.435	1.22–1.69	0.001
							GWAS + Rep	3237	6097	0.045	8.62×10^{-9}	1.523	1.32–1.76	0.002
							GWAS + Rep + Val	5270	8326	0.045	3.18×10^{-7}	1.363	1.21–1.54	0.001
SSc	12q24	<i>ACADS</i>	rs558275	119 681 274	Intergenic	A/G	GWAS	2357	5187	0.398	6.46×10^{-6}	0.848	0.79–0.91	0.457
							GWAS + Rep	3237	6097	0.395	4.99×10^{-6}	0.863	0.81–0.92	0.234
							GWAS + Rep + Val	5270	8326	0.392	8.08×10^{-5}	0.901	0.86–0.95	0.180
ATA +	13q32	<i>IPO5/FAR1</i>	rs586851	97 542 564	Intergenic	C/A	GWAS	462	5187	0.103	1.36×10^{-3}	1.391	1.14–1.70	0.086
							GWAS + Rep	727	6097	0.108	3.89×10^{-6}	1.454	1.24–1.71	0.023
							GWAS + Rep + Val	1161	8326	0.112	4.59×10^{-5}	1.311	1.15–1.49	0.013
SSc	15q24	<i>CSK</i>	rs1378942	72 864 420	Intron	C/A	GWAS	2357	5187	0.362	7.19×10^{-7}	1.199	1.12–1.29	0.573
							GWAS + Rep	3237	6097	0.367	4.42×10^{-8}	1.194	1.12–1.27	0.407
							GWAS + Rep + Val	5270	8326	0.370	5.04×10^{-12}	1.202	1.14–1.27	0.114
ATA +	15q26	<i>ADAMTS17</i>	rs2289584	98 707 019	Intergenic	T/G	GWAS	462	5187	0.175	9.04×10^{-5}	1.391	1.18–1.64	0.383
							GWAS + Rep	727	6097	0.175	6.07×10^{-7}	1.416	1.23–1.62	0.457
							GWAS + Rep + Val	1161	8326	0.178	5.15×10^{-5}	1.257	1.13–1.41	0.099

Chr., chromosome; MAF, minor allele frequency; OR, odds ratio; CI, confidence intervals. GWAS, genome-wide association study; Rep, replication stage (768 SNPs); Val, validation stage (6 SNPs).

^aAll genomic positions are referent to genome build 36.

^bMinor allele first.

^cMantel-Haenszel meta-analysis of the cohorts involved in the corresponding stage of the analysis (see Materials and Methods).

^dAll ORs are for the minor allele.

Table 2. Previously reported associations with SSC found in this study

Chr.	Gene	SNP	Base pair ^a	Location	Change ^b	Sample size Cases	Sample size Controls	MAF	P-value ^c	OR ^d	CI (95%)	Breslow-Day P-value
1q24	<i>CD247</i>	rs2056626	165 687 049	Intron	G/T	3237	6097	0.396	1.14×10^{-8}	0.832	0.78–0.89	5.14×10^{-1}
1q25	<i>TNFSF4</i>	rs10798269	171 576 336	Intergenic	A/G	3237	6097	0.336	1.29×10^{-5}	0.864	0.81–0.92	7.78×10^{-1}
2q32	<i>STAT4</i>	rs4916334	171 600 452	Intergenic	G/T	3237	6097	0.334	1.00×10^{-5}	0.861	0.81–0.92	8.18×10^{-1}
6q23	<i>TNFAIP3</i>	rs10168266 ^e	191 644 049	Intron	T/C	2372	4395	0.214	1.81×10^{-7}	1.234	1.14–1.34	1.11×10^{-1}
7q32	<i>TNPO3/IRF5</i>	rs2230926	138 237 759	Exon	G/T	3237	6097	0.036	2.29×10^{-6}	1.463	1.25–1.71	4.80×10^{-2}
		rs4728142	128 361 203	Intergenic	A/G	3237	6097	0.463	4.74×10^{-10}	1.216	1.14–1.29	9.12×10^{-2}
		rs7808907	128 371 320	Intron	T/C	3237	6097	0.483	1.47×10^{-7}	0.848	0.80–0.90	2.35×10^{-2}
		rs10488631	128 381 419	Intergenic	C/T	3237	6097	0.116	1.58×10^{-18}	1.513	1.38–1.66	6.94×10^{-1}
		rs12531711	128 404 702	Intergenic	G/A	3237	6097	0.115	3.31×10^{-19}	1.527	1.39–1.68	7.11×10^{-1}
		rs12537284	128 505 142	Intergenic	A/G	3237	6097	0.141	1.85×10^{-9}	1.301	1.19–1.42	8.77×10^{-1}
16q24	<i>IRF8</i>	rs2084654	128 516 364	Intergenic	G/A	3237	6097	0.339	1.64×10^{-6}	1.171	1.10–1.25	2.58×10^{-1}
		rs11117425	84 529 772	Intergenic	T/C	3237	6097	0.296	1.43×10^{-6}	0.845	0.79–0.91	1.96×10^{-1}
		rs11644034	84 530 113	Intergenic	A/G	3237	6097	0.196	8.89×10^{-8}	0.804	0.74–0.87	4.03×10^{-1}
		rs12711490	84 530 529	Intergenic	C/T	3237	6097	0.196	6.43×10^{-8}	0.802	0.74–0.87	3.56×10^{-1}
		rs7202472	84 535 003	Intergenic	T/G	3237	6097	0.182	1.58×10^{-7}	0.802	0.74–0.87	3.65×10^{-1}
		rs11642873 ^f	84 549 206	Intergenic	C/A	3237	6097	0.176	2.30×10^{-9}	0.730	0.72–0.86	8.28×10^{-1}

Chr., chromosome; MAF, minor allele frequency; OR, odds ratio; CI, confidence intervals.

^aAll genomic positions are referent to genome build 36.

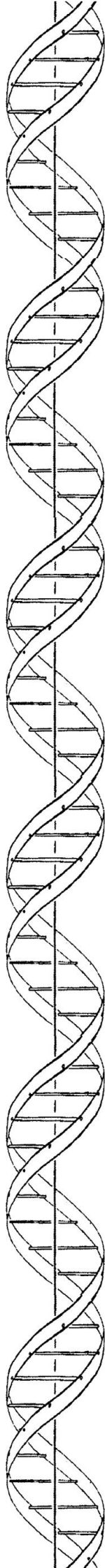
^bMinor allele first.

^cMantel-Haenszel meta-analysis of GWAS and GoldenGate cohorts.

^dAll ORs are for the minor allele.

^e*TAT4* SNP rs10168266 was not genotyped in the European GWAS cohorts, nor could its genotype could be imputed.

^f*IRF8* SNP rs11642873 association was confined to IcSSc and these are the statistical shown in the table rather than global SSC as in the rest of the SNPs.



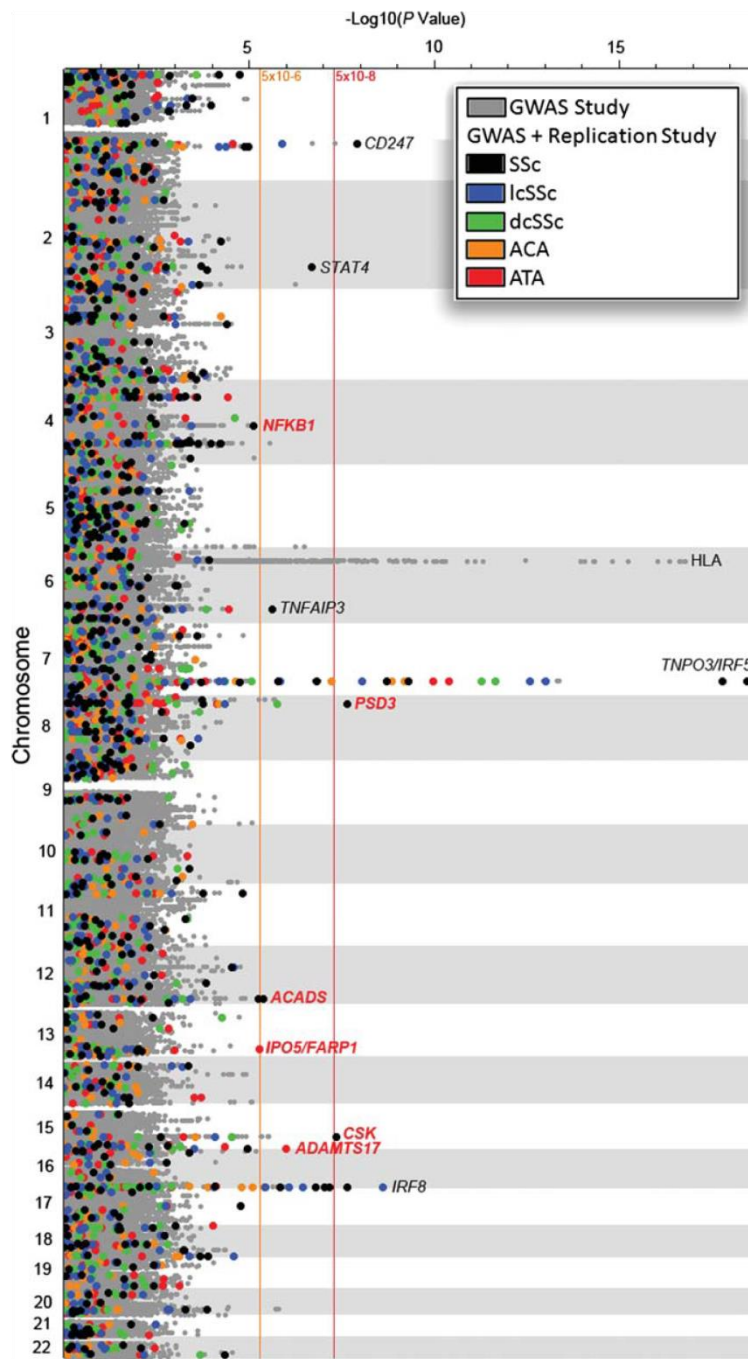


Figure 2. The Manhattan plot of the GWAS and replication cohorts meta-analysis. Grey dots represent P -values of SNPs which are genotyped only in the GWAS cohorts in the previous study by Radstake *et al.* (8). Other color dots represent P_{MH} values of the 720 SNPs which were in both the GWAS and the replication cohorts (Mantel–Haenszel meta-analysis). Gene names in black have been previously reported as genetic risk factors for SSc, while gene names in red are novel ones.

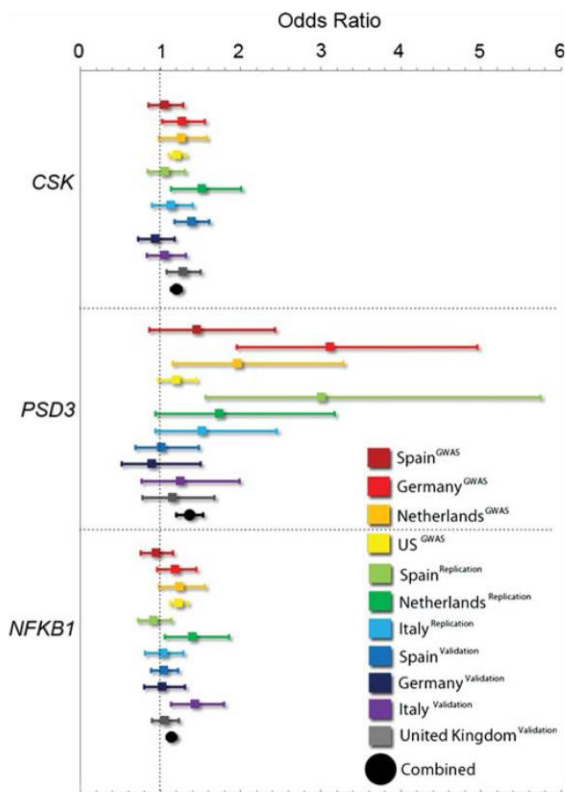


Figure 3. ORs and 95% CIs of each of the novel genetic association with SSc found in this study, in either the meta-analysis of all cohorts and each population separately.

associated with SLE in Asian population (26). These results are nonetheless controversial, since a previous report found no association between this very same in/del and the risk of developing RA or SLE (27). In both the studies, the statistical power was rather limited with cohorts of around 300 patients, hence further fine-mapping studies in the *NFKB1* region with larger study cohorts will be needed to elucidate the role of *NFKB1* genetic variants in autoimmunity in general and SSc specifically.

The *PSD3* gene encodes a protein of unknown function, which has a pleckstrin domain and a Sec7 domain. The pleckstrin domain is found in a wide range of proteins, and is capable of binding phosphatidylinositol, G proteins and protein kinase C, thus acting as a scaffold protein in signal transduction pathways, while the Sec7 domain is a guanine nucleotide exchange factor, which is a component of intracellular signaling networks. Further research will be needed to determine the role of *PSD3*, but these data point to a role in the immune system and the pathogenesis of SSc.

Previously described associations in *CD247* (8), *STAT4* (15,28), *TNFAIP3* (29), *IRF8* (30), *TNFSF4* (31,32) and *TNPO3/IRF5* (33,34) have been firmly replicated in the present study, confirming their role in the pathogenesis of SSc.

Dieude *et al.* (29) found the SNP rs5029939 in *TNFAIP3* associated with SSc, which showed stronger association in the dcSSc and ATA+ subgroups of the disease, although the association remained significant in the opposing lcSSc and ACA+ subgroups. In the present study, rs2230926 (in total LD with rs5029939 in the HapMap CEU population) has shown GWAS level association with the global SSc, and no sign of heterogeneity in the SSc subgroups was detected (Table 2). Taking both studies together, the association of *TNFAIP3* is most likely due to an effect on SSc susceptibility overall rather than any of its sub-phenotypes. The previously reported fluctuating level of significance is probably due to different population composition and cohort size (1656 cases and 1311 controls in the study by Dieude *et al.* (29) and 3237 cases and 6097 controls in the present study). SNP rs2230926 is located in exon 3 of *TNFAIP3*, and encodes for a phenylalanine to cysteine change at residue 127 which has been demonstrated to alter the function of the protein, which makes it a functionally associated genetic variant with SLE (35).

Association in the *TNPO3/IRF5* region has been detected previously in SSc (8,33,34) and SLE (36,37). The three independently associated polymorphisms have been detected in SLE: an in/del in exon 6 which change the expression level of *IRF5* (tagged by rs10488631 in our study), a variant in exon 1B which affects splicing of this exon (not tagged in our study) and a variant which disrupts a polyA signal site (tagged by rs4728142 in our study). Our findings in the present study are consistent with associations found in SLE. We detect two of the three independent variants (Table 3): the in/del and the polyA signal, yet the variant in exon 1B was not captured in this study. This observation adds more evidence for a similar genetic component of SSc and SLE. Nevertheless, further research will be needed in order to determine whether the variant in exon 1B, affecting splicing of the gene, also has a major role in SSc.

IRF8 has been recently described as a risk factor for lcSSc (30) and has also been associated with the risk of multiple sclerosis (38). Of all the SNPs selected for replication in the present study, only rs11642873 remained independently significant (Table 3), and its association was found, as in the previous study, exclusively confined to lcSSc. Further fine-mapping and functional studies are necessary in order to determine the role of genetic variants in *IRF8* in the pathogenesis of SSc and autoimmunity.

In summary, by the analysis of 720 genetic variants, selected from GWAS data's 'grey zone' of association, we describe three new genetic risk factors for SSc (*CSK*, *PSD3* and *NFKB1*) and confirm five previously reported associations (*CD247*, *STAT4*, *TNFAIP3*, *TNPO3/IRF5*, *TNFSF4* and *IRF8*). Also in the case of *TNPO3/IRF5*, we clarify the nature of the association with SSc which is similar to that found in SLE, thus highlighting the similarities of the genetic component for both diseases.

MATERIALS AND METHODS

Populations

The four populations used in the previous SSc GWAS have been previously described (8). Replication and validation

Table 3. Conditional analysis in the TNPO3/IRF5 and IRF8 genetic regions

Gene	Chr.	SNP	Base pair ^a	Location	Change ^b	MAF	P_{MH}	OR ^c	CI (95%)	Breslow–Day P -value	Conditioned P -value
TNPO3/IRF5	7q32	rs4728142	128 361 203	Intergenic	A/G	0.463	4.74E – 10	1.216	1.14–1.29	9.12E – 02	9.86 × 10⁻⁴
		rs7808907	128 371 320	Intron	T/C	0.483	1.47E – 07	0.848	0.80–0.90	2.35E – 02	9.41 × 10 ⁻¹
		rs10488631	128 381 419	Intergenic	C/T	0.116	1.58E – 18	1.513	1.38–1.66	6.94E – 01	1.96 × 10⁻⁷
		rs12531711	128 404 702	Intergenic	G/A	0.115	3.31E – 19	1.527	1.39–1.68	7.11E – 01	2.46 × 10⁻⁷
		rs12537284	128 505 142	Intergenic	A/G	0.141	1.85E – 09	1.301	1.19–1.42	8.77E – 01	2.43 × 10 ⁻¹
IRF8	16q24	rs2084654	128 516 364	Intergenic	G/A	0.339	1.64E – 06	1.171	1.10–1.25	2.58E – 01	1.36 × 10 ⁻¹
		rs8056420	83 969 142	Intergenic	G/A	0.162	5.49 × 10 ⁻¹	1.029	0.94–1.13	3.13 × 10 ⁻³	8.40 × 10 ⁻¹
		rs7186021	84 043 560	Intergenic	C/A	0.473	2.31 × 10 ⁻²	0.922	0.86–0.99	2.48 × 10 ⁻³	5.32 × 10 ⁻¹
		rs8053194	84 072 623	Intergenic	T/G	0.404	3.01 × 10 ⁻²	0.925	0.86–0.99	6.69 × 10 ⁻³	6.92 × 10 ⁻¹
		rs1117425	84 529 772	Intergenic	T/C	0.296	1.43E – 06	0.845	0.79–0.90	1.96E – 01	6.85 × 10 ⁻²
		rs11644034	84 530 113	Intergenic	A/G	0.196	8.89E – 08	0.8043	0.74–0.87	4.03E – 01	3.68 × 10 ⁻¹
		rs12711490	84 530 529	Intergenic	C/T	0.196	6.43E – 08	0.802	0.74–0.87	3.56E – 01	7.90 × 10 ⁻¹
		rs7202472	84 535 003	Intergenic	T/G	0.182	1.58E – 07	0.802	0.74–0.87	3.65E – 01	6.07 × 10 ⁻¹
		rs11642873	84 549 206	Intergenic	C/A	0.179	2.30E – 09	0.730	0.66–0.81	9.21E – 01	2.05 × 10⁻⁵
		rs10514613	84 688 032	Intergenic	C/T	0.056	4.81 × 10 ⁻²	1.162	1.00–1.35	6.44 × 10 ⁻⁴	3.34 × 10 ⁻²

SNPs in bold form a haplotype.

Chr., chromosome; MAF, minor allele frequency; OR, odds ratio; CI, confidence intervals.

^aAll genomic positions are referent to genome build 36.

^bMinor allele first.

^cAll ORs are for the minor allele.

cohorts from Spain (two independent cohorts), Germany, Netherlands, Italy (two independent cohorts) and the UK for a total of 2913 cases and 3139 controls were recruited from hospitals and blood banks from each of these countries (numbers before quality control filters). Key features of these populations can be found in Supplementary Material, Table S1. This study was approved by the local ethics committees of the participating hospitals. All of the SSc patients participating in this study met American College of Rheumatology criteria (39) and were classified in the disease subtypes according to LeRoy and Medsger (10). All individuals in this study were of European ancestry [either self-reported and/or principal component analysis (PCA) determined, see below] and gave written informed consent.

Samples in this study included those in the initial discovery GWAS cohorts as well as those in the replication cohorts in the work by Radstake *et al.* (8). In the present study, the previous replication cohorts were genotyped for 768 additional SNPs.

Combination of GWAS and replication cohorts genotyped in this study resulted in a total of 5270 cases and 8326 controls (after QC was applied), which gave us a statistical power of 100% to detect an OR of 1.3 with a minor allele frequency of 0.20. Other power calculations can be found in Supplementary Material, Table S2.

Study design

SNP selection. First, we aimed to select the putative SSc-associated genetic variants that did not reach genome-wide significance in our initial GWAS (8). We filtered GWAS data as previously published. We excluded from our analysis the extended HLA region on chromosome six, since association of this region with SSc is well known, and was not the focus of the present study. Since there is growing evidence for the existence of genetic heterogeneity in SSc, we

included the clinical subtypes of the disease (i.e. lcSSc and dcSSc) and the presence of the two most common auto-antibodies (i.e. ACA and ATA) in our selection criteria (1,3,10). Because of the differences between the US and European samples in the PCA, we decided to select the SNPs for each set separately, lending greater weight to the European cohorts (it should be noted that all our replication cohorts were of European countries).

We selected all SNPs with a Mantel–Haenszel, λ corrected P -value lower than 10^{-3} from the European cohorts (i.e. Netherlands, Germany and Spain) meta-analysis. We also selected all SNPs with a Mantel–Haenszel λ -corrected P -value lower than 10^{-4} in the lcSSc, dcSSc, ACA+ and ATA+ subgroups. We also considered the possibility that SNPs associated in one of the three populations could have been filtered out in the others due to any criteria, so we included all SNPs with a corrected P -value lower than 10^{-4} in any of the three populations before merging the data sets. This gave a total of 676 SNPs selected. In order to use all the genotyping capacity of the chosen platform (in which 768 SNPs could be analyzed), and taking into consideration that the first 676 SNPs were selected only from the European panel, we also included the 92 previously non-described independent signals from the US panel that showed the most significant P -values. The overall strategy followed in the present study can be found in Figure 1.

Replication stage. In this stage, we aimed to confirm selected associations from GWAS data on case/control cohorts from Spain, Holland and Italy up to a total of 880 cases and 910 controls. All individuals in these cohorts were genotyped for the 768 selected SNPs and association analyses were performed.

Validation stage. At this point, we selected the top SNPs from the meta-analysis of the previous stages (those genotyped in

GWAS and GoldenGate platforms) and further tested their association in larger cohorts from Spain, Germany, Italy and the UK up to a total of 2033 cases and 2229 controls. All SNPs with a meta-analysis P -value $< 1 \times 10^{-5}$ either in the disease or any of its subphenotypes (i.e. lcSSc, dcSSc, ACA+ or ATA+) and not previously reported to be associated with SSc were selected for validation, comprising a total of six SNPs. Finally, we performed meta-analyses for all associated SNPs combining the GWAS and all replication cohorts.

Genotyping methods

The replication cohort was genotyped for 768 SNPs, selected from the GWAS analysis as described above, using a custom Illumina GoldenGate array run on the Illumina iScan system. The validation stage was genotyped using TaqMan genotyping assays from Applied Biosystems.

Genotype imputation of data

As described in Radstake *et al.* (8), the different cohorts were genotyped using different genotyping arrays (the European cohorts were genotyped mostly with the Illumina HumanHap 370k array, whereas the US cohort was genotyped with the Illumina 550k array). As a consequence, it was possible that some SNPs were present in one but not in the other platform. To prevent this, a genotype imputation was performed in all the GWAS cohorts to obtain a full overlap between platforms. Imputation was performed with IMPUTE software 1.00 as previously described (40), using as reference panels the CEU and TSI HapMap populations.

Data analysis

All data were analyzed using Plink software version 1.07 (<http://pngu.mgh.harvard.edu/purcell/plink/>) (41). LD patterns among SNPs were also calculated with the r^2 statistic using HaploView software (42). Manhattan plots were generated using HelixTree SNP Variation Suite 7 (http://www.goldenhelix.com/SNP_Variation/HelixTree/index.html). In order to avoid any position discrepancies, all base-pair locations for the analyzed genetic variations correspond to those reported in the genome build 36, as these were the positions used in the original GWAS.

All data were quality filtered using the following criteria: Hardy–Weinberg equilibrium P -value > 0.001 on controls of any of the populations analyzed separately, minor allele frequency > 0.01 , success call rate per individual > 0.95 and per SNP > 0.95 . After applying quality control, remaining SNPs were statistically analyzed using the Chi-squared case/control approach in 1433 cases and 1644 controls. The meta-analysis of the different cohorts was conducted using the Mantel–Haenszel test to calculate a pooled OR, and the 95% confidence interval for the OR was estimated using a random effect model. Heterogeneity between cohorts was tested using the Breslow–Day test (P -values < 0.05 were considered statistically significant); nevertheless, we did not rule out any association in the basis of OR heterogeneity. Combined P -values for the Mantel–Haenszel tests were calculated

as implemented in Plink. The independence of effects of SNPs in the same genetic region was tested by multiple logistic regression analysis, conditioning each SNP to all others, as implemented in the Plink software. Population stratification was assessed by PCA as previously described (43). All individuals who deviated more than four standard deviations from the centroid of their population were removed as outliers.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

We thank Sofia Vargas, Sonia Garcia and Gema Robledo for their excellent technical assistance, and all the patients and healthy controls for kindly accepting their essential collaboration. We would also like to thank the following organizations: the EULAR Scleroderma Trials and Research (EUSTAR), the German Network of Systemic Sclerosis and Banco Nacional de ADN (University of Salamanca, Spain).

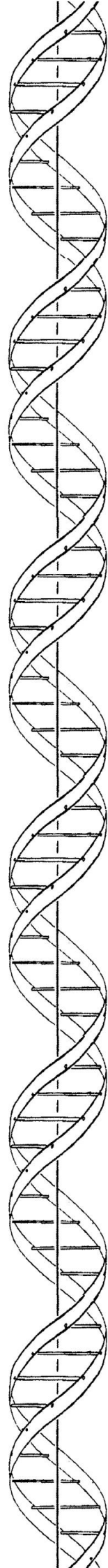
Conflict of Interest statement: None declared.

FUNDING

This work was supported by the following grants: J.M. was funded by GEN-FER from the Spanish Society of Rheumatology, SAF2009-11110 from the Spanish Ministry of Science, CTS-4977 from Junta de Andalucía (Spain), Redes Temáticas de Investigación Cooperativa Sanitaria Program, RD08/0075 (RIER) from Instituto de Salud Carlos III (ISCIII, Spain) and Fondo Europeo de Desarrollo Regional (FEDER). T.R.D.J.R. was funded by the VIDI laureate from the Dutch Association of Research (NWO) and Dutch Arthritis Foundation (National Reumafonds). J.M. and T.R.D.J.R. were sponsored by the Orphan Disease Program grant from the European League Against Rheumatism (EULAR). B.P.C.K. is supported by the Dutch Diabetes Research Foundation (grant 2008.40.001) and the Dutch Arthritis Foundation (Reumafonds, grant NR 09-1-408). T.W. is supported by the grant DFG WI 1031/6-1 and by the grant DFG WI 1031/6-1. N.O.-C. was funded by PI-0590-2010, Consejería de Salud, Junta de Andalucía, Spain. F.K.T., F.C.A. and M.D.M. were supported by NIH Scleroderma Family Registry and DNA Repository (N01-AR-0-2251), NIH RO1-AR055258 and NIH Center of Research Translation in Scleroderma (1P50AR054144), and the Department of Defense Congressionally Directed Medical Research Programs (W81XWH-07-01-0111). C.F. was supported by 'The Raynaud's and Scleroderma Association' and 'The Scleroderma Society'.

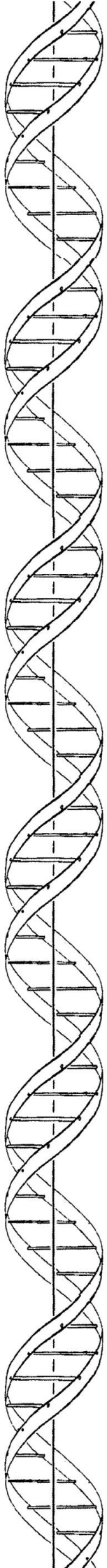
REFERENCES

- Gabrielli, A., Avvedimento, E.V. and Krieg, T. (2009) Scleroderma. *N. Engl. J. Med.*, **360**, 1989–2003.
- Agarwal, S.K., Tan, F.K. and Arnett, F.C. (2008) Genetics and genomic studies in scleroderma (systemic sclerosis). *Rheum. Dis. Clin. North Am.*, **34**, 17–40.



3. Jimenez, S.A. and Derk, C.T. (2004) Following the molecular pathways toward an understanding of the pathogenesis of systemic sclerosis. *Ann. Intern. Med.*, **140**, 37–50.
4. Romano, E., Manetti, M., Guiducci, S., Ceccarelli, C., Allanore, Y. and Matucci-Cerinic, M. (2011) The genetics of systemic sclerosis: an update. *Clin. Exp. Rheumatol.*, **29**, S75–S86.
5. Martin, J. and Fonseca, C. (2011) The genetics of scleroderma. *Curr. Rheumatol. Rep.*, **13**, 13–20.
6. Khor, B., Gardet, A. and Xavier, R.J. (2011) Genetics and pathogenesis of inflammatory bowel disease. *Nature*, **474**, 307–317.
7. Sestak, A.L., Furnrohr, B.G., Harley, J.B., Merrill, J.T. and Namjou, B. (2011) The genetics of systemic lupus erythematosus and implications for targeted therapy. *Ann. Rheum. Dis.*, **70**(Suppl. 1), 37–43.
8. Radstake, T.R., Gorlova, O., Rueda, B., Martin, J.E., Alizadeh, B.Z., Palomino-Morales, R., Coenen, M.J., Vonk, M.C., Voskuyl, A.E., Schuerwegh, A.J. et al. (2010) Genome-wide association study of systemic sclerosis identifies CD247 as a new susceptibility locus. *Nat. Genet.*, **42**, 426–429.
9. Allanore, Y., Saad, M., Dieude, P., Avouac, J., Distler, J.H., Amouyel, P., Matucci-Cerinic, M., Riemekasten, G., Airo, P., Melchers, I. et al. (2011) Genome-wide scan identifies TNIP1, PSORS1C1, and RHOB as novel risk loci for systemic sclerosis. *PLoS Genet.*, **7**, e1002091.
10. LeRoy, E.C. and Medsger, T.A. Jr (2001) Criteria for the classification of early systemic sclerosis. *J. Rheumatol.*, **28**, 1573–1576.
11. Nietert, P.J., Mitchell, H.C., Bolster, M.B., Shaftman, S.R., Tilley, B.C. and Silver, R.M. (2006) Racial variation in clinical and immunological manifestations of systemic sclerosis. *J. Rheumatol.*, **33**, 263–268.
12. Assassi, S., Arnett, F.C., Reveille, J.D., Gourh, P. and Mayes, M.D. (2007) Clinical, immunologic, and genetic features of familial systemic sclerosis. *Arthritis Rheum.*, **56**, 2031–2037.
13. Steen, V.D. (2008) The many faces of scleroderma. *Rheum. Dis. Clin. North Am.*, **34**, 1–15. ; v.
14. Arnett, F.C., Gourh, P., Shete, S., Ahn, C.W., Honey, R.E., Agarwal, S.K., Tan, F.K., McNearney, T., Fischbach, M., Fritzler, M.J. et al. (2010) Major histocompatibility complex (MHC) class II alleles, haplotypes and epitopes which confer susceptibility or protection in systemic sclerosis: analyses in 1300 Caucasian, African-American and Hispanic cases and 1000 controls. *Ann. Rheum. Dis.*, **69**, 822–827.
15. Rueda, B., Broen, J., Simeon, C., Hesselstrand, R., Diaz, B., Suarez, H., Ortego-Centeno, N., Riemekasten, G., Fonollosa, V., Vonk, M.C. et al. (2009) The STAT4 gene influences the genetic predisposition to systemic sclerosis phenotype. *Hum. Mol. Genet.*, **18**, 2071–2077.
16. Rueda, B., Gourh, P., Broen, J., Agarwal, S.K., Simeon, C., Ortego-Centeno, N., Vonk, M.C., Coenen, M., Riemekasten, G., Hunzelmann, N. et al. (2010) BANK1 functional variants are associated with susceptibility to diffuse systemic sclerosis in Caucasians. *Ann. Rheum. Dis.*, **69**, 700–705.
17. Gourh, P., Tan, F.K., Assassi, S., Ahn, C.W., McNearney, T.A., Fischbach, M., Arnett, F.C. and Mayes, M.D. (2006) Association of the PTPN22 R620W polymorphism with anti-topoisomerase I- and anticentromere antibody-positive systemic sclerosis. *Arthritis Rheum.*, **54**, 3945–3953.
18. Lowell, C.A. (2004) Src-family kinases: rheostats of immune cell signaling. *Mol. Immunol.*, **41**, 631–643.
19. Okutani, D., Lodyga, M., Han, B. and Liu, M. (2006) Src protein tyrosine kinase family and acute inflammatory responses. *Am. J. Physiol. Lung Cell Mol. Physiol.*, **291**, 129–141.
20. Vittal, R., Horowitz, J.C., Moore, B.B., Zhang, H., Martinez, F.J., Toews, G.B., Standiford, T.J. and Thannickal, V.J. (2005) Modulation of pro-survival signaling in fibroblasts by a protein kinase inhibitor protects against fibrotic tissue injury. *Am. J. Pathol.*, **166**, 367–375.
21. Thannickal, V.J., Lee, D.Y., White, E.S., Cui, Z., Larios, J.M., Chacon, R., Horowitz, J.C., Day, R.M. and Thomas, P.E. (2003) Myofibroblast differentiation by transforming growth factor-beta1 is dependent on cell adhesion and integrin signaling via focal adhesion kinase. *J. Biol. Chem.*, **278**, 12384–12389.
22. Skhirtladze, C., Distler, O., Dees, C., Akhmetshina, A., Busch, N., Venalis, P., Zwerina, J., Spriewald, B., Pilecky, M., Schett, G. et al. (2008) Src kinases in systemic sclerosis: central roles in fibroblast activation and in skin fibrosis. *Arthritis Rheum.*, **58**, 1475–1484.
23. Trynka, G., Hunt, K.A., Bockett, N.A., Romanos, J., Mistry, V., Szperl, A., Bakker, S.F., Bardella, M.T., Bhaw-Rosun, L., Castillejo, G. et al. (2011) Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat. Genet.*, **43**, 1193–1201.
24. Barnes, P.J. and Karin, M. (1997) Nuclear factor-kappaB: a pivotal transcription factor in chronic inflammatory diseases. *N. Engl. J. Med.*, **336**, 1066–1071.
25. Tak, P.P. and Firestein, G.S. (2001) NF-kappaB: a key role in inflammatory diseases. *J. Clin. Invest.*, **107**, 7–11.
26. Gao, M., Wang, C.H., Sima, X. and Han, X.M. (2011) NFKB1 -94 insertion/deletion ATTG polymorphism contributes to risk of systemic lupus erythematosus. *DNA Cell Biol.* [Epub ahead of print].
27. Orozco, G., Sanchez, E., Collado, M.D., Lopez-Nevot, M.A., Paco, L., Garcia, A., Jimenez-Alonso, J. and Martin, J. (2005) Analysis of the functional NFKB1 promoter polymorphism in rheumatoid arthritis and systemic lupus erythematosus. *Tissue Antigens*, **65**, 183–186.
28. Dieude, P., Guedj, M., Wipff, J., Ruiz, B., Hachulla, E., Diot, E., Granel, B., Sibilia, J., Tiev, K., Mouthon, L. et al. (2009) STAT4 is a genetic risk factor for systemic sclerosis having additive effects with IRF5 on disease susceptibility and related pulmonary fibrosis. *Arthritis Rheum.*, **60**, 2472–2479.
29. Dieude, P., Guedj, M., Wipff, J., Ruiz, B., Riemekasten, G., Matucci-Cerinic, M., Melchers, I., Hachulla, E., Airo, P., Diot, E. et al. (2010) Association of the TNFAIP3 rs5029939 variant with systemic sclerosis in the European Caucasian population. *Ann. Rheum. Dis.*, **69**, 1958–1964.
30. Gorlova, O., Martin, J.E., Rueda, B., Koeleman, B.P., Ying, J., Teruel, M., Diaz-Gallo, L.M., Broen, J.C., Vonk, M.C., Simeon, C.P. et al. (2011) Identification of novel genetic markers associated with clinical phenotypes of systemic sclerosis through a genome-wide association strategy. *PLoS Genet.*, **7**, e1002178.
31. Bossini-Castillo, L., Broen, J.C., Simeon, C.P., Beretta, L., Vonk, M.C., Ortego-Centeno, N., Espinosa, G., Carreira, P., Camps, M.T., Navarrete, N. et al. (2011) A replication study confirms the association of TNFSF4 (OX40L) polymorphisms with systemic sclerosis in a large European cohort. *Ann. Rheum. Dis.*, **70**, 638–641.
32. Gourh, P., Arnett, F.C., Tan, F.K., Assassi, S., Divecha, D., Paz, G., McNearney, T., Draeger, H., Reveille, J.D., Mayes, M.D. et al. (2010) Association of TNFSF4 (OX40L) polymorphisms with susceptibility to systemic sclerosis. *Ann. Rheum. Dis.*, **69**, 550–555.
33. Dieude, P., Guedj, M., Wipff, J., Avouac, J., Fajardy, I., Diot, E., Granel, B., Sibilia, J., Cabane, J., Mouthon, L. et al. (2009) Association between the IRF5 rs2004640 functional polymorphism and systemic sclerosis: a new perspective for pulmonary fibrosis. *Arthritis Rheum.*, **60**, 225–233.
34. Ito, I., Kawaguchi, Y., Kawasaki, A., Hasegawa, M., Ohashi, J., Hikami, K., Kawamoto, M., Fujimoto, M., Takehara, K., Sato, S. et al. (2009) Association of a functional polymorphism in the IRF5 region with systemic sclerosis in a Japanese population. *Arthritis Rheum.*, **60**, 1845–1850.
35. Musone, S.L., Taylor, K.E., Lu, T.T., Nititham, J., Ferreira, R.C., Ortmann, W., Shifrin, N., Petri, M.A., Kambh, M.I., Manzi, S. et al. (2008) Multiple polymorphisms in the TNFAIP3 region are independently associated with systemic lupus erythematosus. *Nat. Genet.*, **40**, 1062–1064.
36. Sigurdsson, S., Nordmark, G., Goring, H.H., Lindroos, K., Wiman, A.C., Sturfelt, G., Jonsen, A., Rantapaa-Dahlqvist, S., Moller, B., Kere, J. et al. (2005) Polymorphisms in the tyrosine kinase 2 and interferon regulatory factor 5 genes are associated with systemic lupus erythematosus. *Am. J. Hum. Genet.*, **76**, 528–537.
37. Graham, R.R., Kyogoku, C., Sigurdsson, S., Vlasova, I.A., Davies, L.R., Baechler, E.C., Plenge, R.M., Koeuth, T., Ortmann, W.A., Hom, G. et al. (2007) Three functional variants of IFN regulatory factor 5 (IRF5) define risk and protective haplotypes for human lupus. *Proc. Natl Acad. Sci. USA*, **104**, 6758–6763.
38. De Jager, P.L., Jia, X., Wang, J., de Bakker, P.I., Ottoboni, L., Aggarwal, N.T., Piccio, L., Raychaudhuri, S., Tran, D., Aubin, C. et al. (2009) Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nat. Genet.*, **41**, 776–782.
39. 1980) Preliminary criteria for the classification of systemic sclerosis (scleroderma). Subcommittee for scleroderma criteria of the American Rheumatism Association Diagnostic and Therapeutic Criteria Committee. *Arthritis Rheum.*, **23**, 581–590.

40. Marchini, J., Howie, B., Myers, S., McVean, G. and Donnelly, P. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.*, **39**, 906–913.
41. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
42. Barrett, J.C., Fry, B., Maller, J. and Daly, M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.
43. Tian, C., Kosoy, R., Nassir, R., Lee, A., Villoslada, P., Klareskog, L., Hammarstrom, L., Garchon, H.J., Pulver, A.E., Ransom, M. *et al.* (2009) European population genetic substructure: further definition of ancestry informative markers for distinguishing among diverse European ethnic groups. *Mol. Med.*, **15**, 371–383.

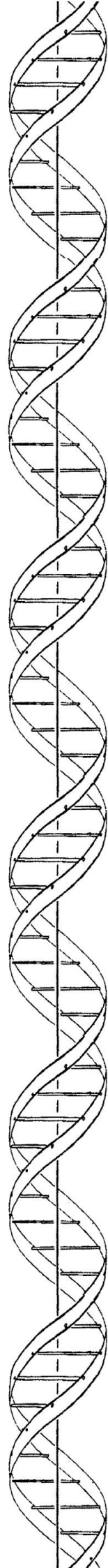


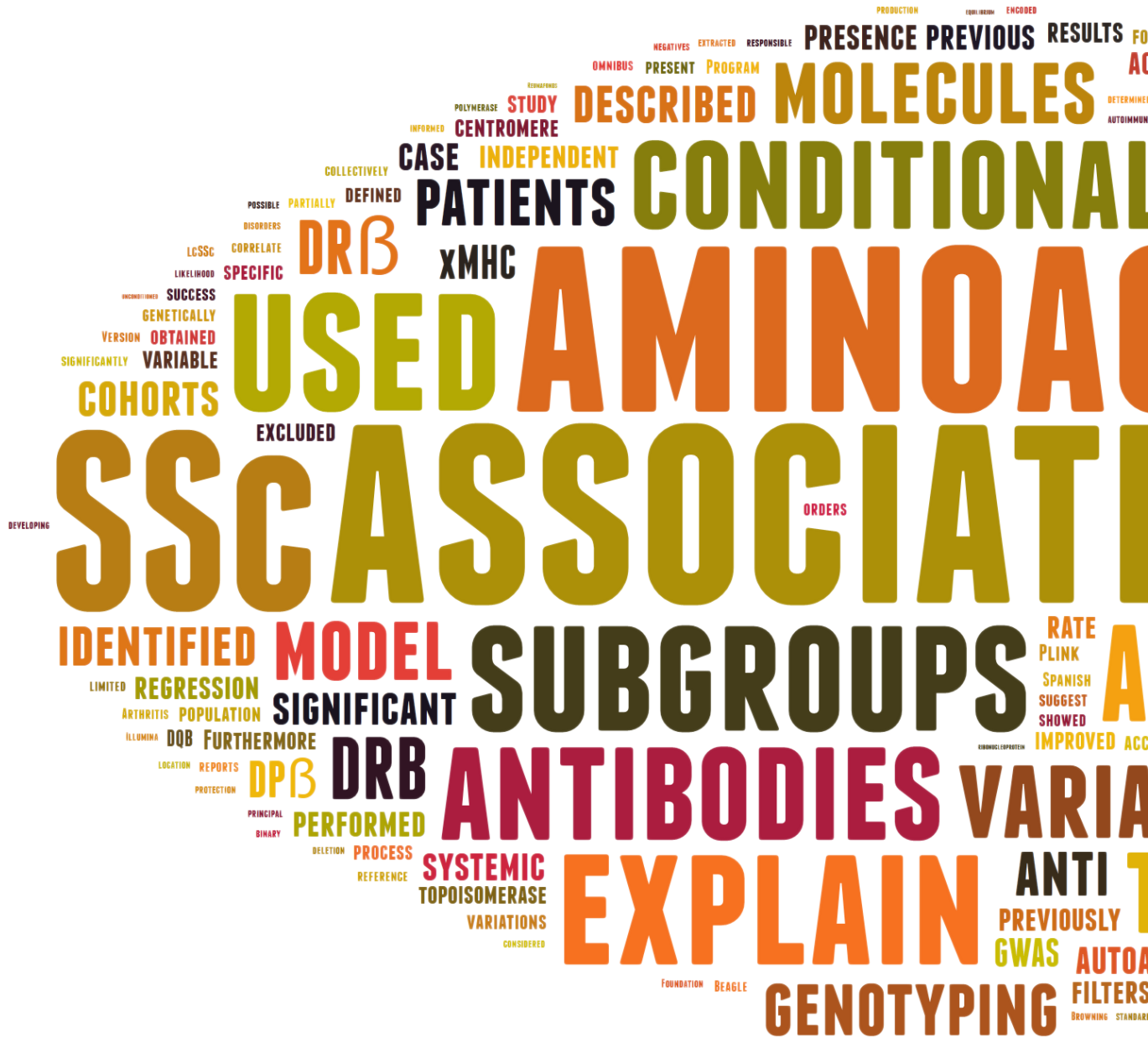
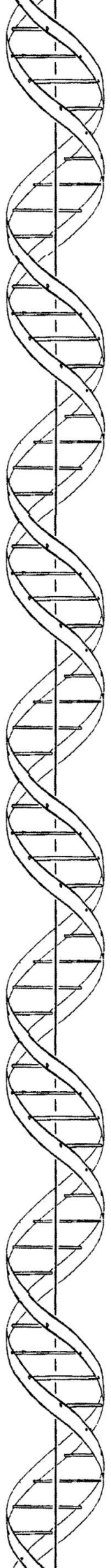
Supplementary table 1. Composition, subtype and auto-antibody status of each population included in the study.

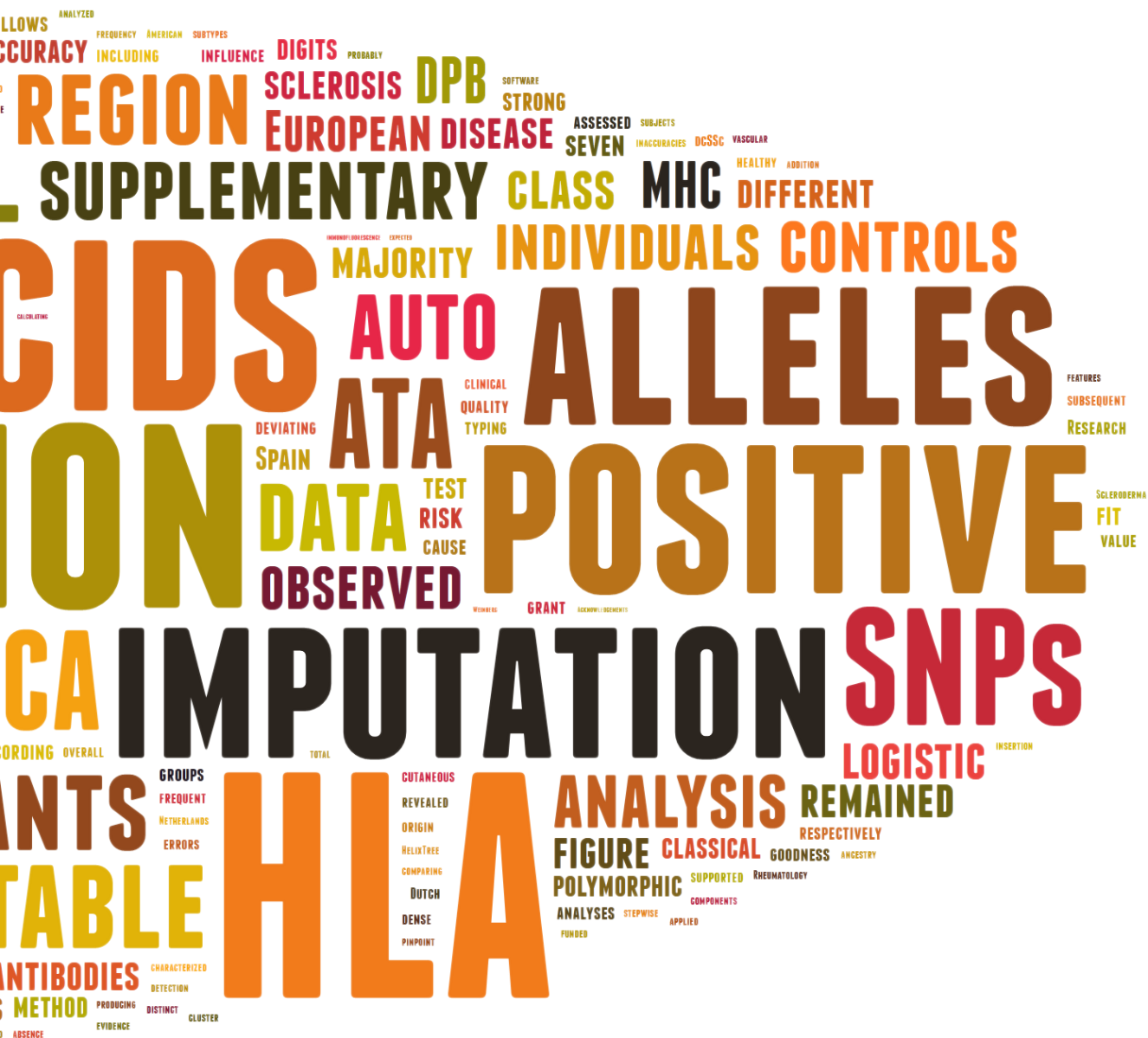
Population	Population Size		Gender (cases/controls)		Subtype		ACA Positive	ATA Positive
	Cases	Controls	Female	Male	Diffuse	Limited		
Overall	5,270	8,326	0.86/0.75	0.14/0.25	0.33	0.67	0.37	0.24
GWAS SSc cohorts								
Spain	376	389	0.90/0.73	0.10/0.27	0.30	0.70	0.50	0.26
Germany	286	670	0.88/0.62	0.12/0.38	0.43	0.57	0.45	0.33
The Netherlands	203	643	0.72/0.51	0.28/0.49	0.24	0.76	0.25	0.28
US	1,492	3,485	0.88/0.88	0.12/0.12	0.36	0.64	0.32	0.17
Replication SSc cohorts								
Spain	314	348	0.91/0.37	0.09/0.63	0.39	0.61	0.45	0.27
The Netherlands	195	240	0.71/0.47	0.29/0.53	0.39	0.61	0.25	0.28
Italy	371	322	0.91/0.51	0.09/0.49	0.25	0.75	0.43	0.38
Validation SSc cohorts								
Spain	527	1,016	0.84/0.66	0.16/0.34	0.30	0.70	0.37	0.23
Germany	284	289	0.80/0.52	0.20/0.48	0.41	0.59	0.34	0.33
Italy	306	372	0.93/0.66	0.07/0.34	0.28	0.72	0.51	0.30
UK	916	552	0.83/0.81	0.17/0.19	0.29	0.71	0.36	0.18

Supplementary table 2. Power calculations for the present study cohort size in different scenarios based in expected odds ratio and minor allele frequency.

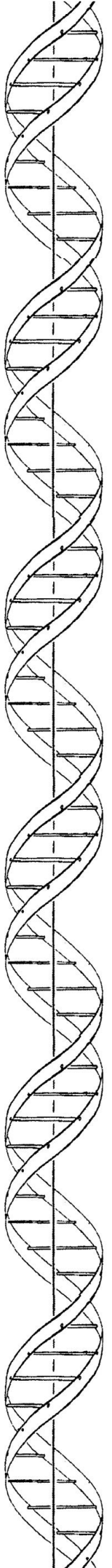
Phenotype	N		Level of Significance	OR 1.30						OR1.20					
	Cases	Controls		MAF	MAF	MAF	MAF	MAF	MAF	MAF	MAF	MAF	MAF	MAF	
				0.40	0.30	0.20	0.15	0.10	0.05	0.40	0.30	0.20	0.15	0.10	0.05
GWAS	2,357	5,187	5x10-8	98	95	77	54	21	2	37	26	11	4	1	0
Replication	880	910	5x10-8	6	4	2	1	0	0	0	0	0	0	0	0
GWAS+Replication	3,237	6,097	5x10-8	100	100	95	82	46	5	66	52	27	13	3	0
Validation	2,033	2,229	5x10-8	73	61	35	18	5	0	10	6	2	1	0	0
Joint	5,270	8,326	5x10-8	100	100	100	99	89	27	96	91	71	47	17	1

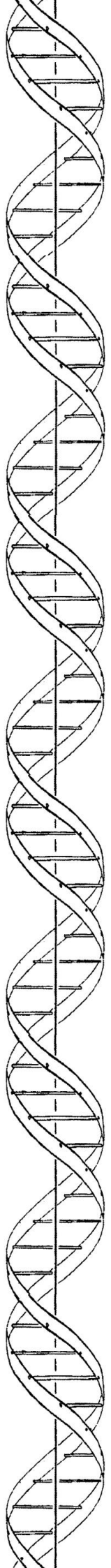






SEVEN AMINOACIDS IN *HLA-DRB1* AND *HLA-DPB1*
EXPLAIN THE MAJORITY OF MHC ASSOCIATIONS WITH
SYSTEMIC SCLEROSIS. UNDER REVIEW.





Seven aminoacids in HLA-DRB1 and HLA-DPB1 explain the majority of MHC associations with systemic sclerosis

Jose-Ezequiel Martin^{1,†,*}, Paul I.W. de Bakker^{2,†}, Carmen P. Simeon³, Norberto Ortego-Centeno⁴, Patricia Carreira⁵, Miguel A. Gonzalez-Gay⁶, Nicolas Hunzelmann⁷, Madelon C. Vonk⁸, Annemie J. Schuerwegh⁹, Alexandre E. Voskuyl¹⁰, Gabriela Riemekasten^{11,12}, Torsten Witte¹³, Olga Gorlova¹⁴, Frank C. Arnett¹⁵, Xiaodong Zhou¹⁵, Shervin Assassi¹⁵, John D. Reveille¹⁵, Timothy R.D.J. Radstake^{16,†}, Maureen D. Mayes^{15,†}, Javier Martin^{1,†}, Bobby P.C. Koeleman^{2,†}.

Affiliations

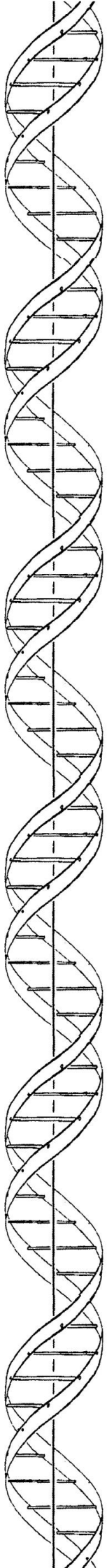
¹Instituto de Parasitología y Biomedicina López-Neyra, CSIC, Granada, Spain.

²Department of Medical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands. ³Servicio de Medicina Interna, Hospital Valle de Hebron, Barcelona, Spain.

⁴Servicio de Medicina Interna, Hospital Clínico Universitario, Granada, Spain. ⁵Hospital 12 de Octubre, Madrid, Spain. ⁶Servicio de Reumatología, Hospital Marqués de Valdecilla, Santander, Spain. ⁷Department of Dermatology, University of Cologne, Cologne, Germany.

⁸Department of Rheumatology, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands. ⁹Department of Rheumatology, University of Leiden, Leiden, The Netherlands. ¹⁰VU University Medical Center, Amsterdam, The Netherlands. ¹¹Department of Rheumatology and Clinical Immunology, Charité University Hospital, Berlin, Germany.

¹²German Rheumatism Research Centre, Leibniz Institute, Germany. ¹³Hannover Medical School, Hannover, Germany. ¹⁴Department of Epidemiology, M.D. Anderson Cancer Center, Houston, Texas, USA. ¹⁵The University of Texas Health Science Center–Houston,



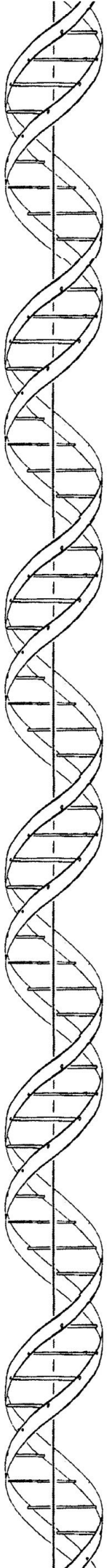
Houston, TX, USA. ¹⁶Department of Rheumatology and Clinical Immunology, Utrecht Medical Center, The Netherlands. [†]These authors contributed equally to this work.

*To whom attention correspondence should be written: cebercoto@ipb.csic.es.

Keywords: HLA, MHC, systemic sclerosis, scleroderma, genetic risk, imputation, autoimmunity.

Abstract

Distinct HLA alleles are associated with systemic sclerosis (SSc), but they collectively do not explain the strong association signal observed for the HLA region in recent genome wide association studies. Here, we took advantage of existing dense genotype data and imputed the HLA class I and II alleles, together with 894 polymorphic aminoacidic positions and 3,841 SNPs, in 2,296 cases and 5,356 controls of European origin. Conditional analyses revealed distinct signatures of association within SSc subtypes related to two auto-antibodies anti-centromere (ACA) and anti-topoisomerase I (ATA) status. Three variable aminoacids in positions 13, 60 and 71 of the HLA-DR β 1 molecule explain the majority of HLA association with ACA. Similarly, variable aminoacids at position 76 of the HLA-DP β 1 molecule and 58, 67 and 86 of the HLA-DR β 1 molecule explain the majority of association towards ATA production. No significant association remains after controlling for these two groups of aminoacids. These results suggest that the HLA association with SSc is different between the autoantibody subgroups and is determined by the two groups of aminoacids.

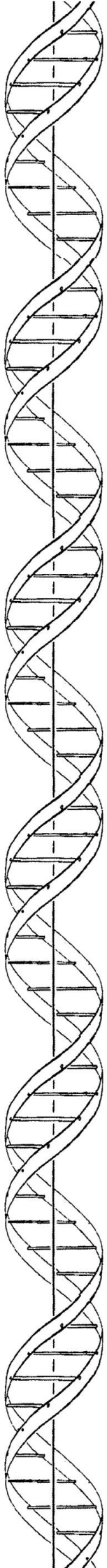


Author Summary

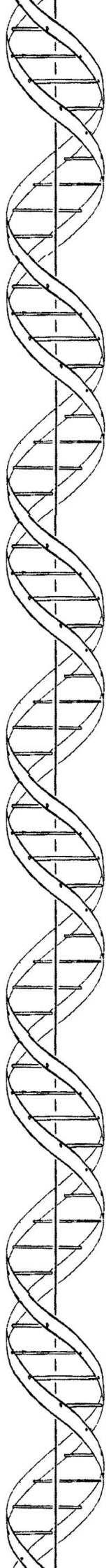
Different HLA alleles have been largely known to influence the risk of developing autoimmune disorders, among which we find systemic sclerosis, a chronic connective tissue disorder characterized by an abnormal immune process, excessive collagen deposit and vascular damage. Nevertheless, the specific variations responsible for the observed association in the HLA region with systemic sclerosis have remained elusive. In this study we determine, through a novel HLA imputation technique, the specific aminoacids in the HLA molecules which cause an increased risk of developing systemic sclerosis and its associated auto-antibodies. With a seven aminoacid model (in the HLA-DR β 1 and HLA-DP β 1 molecules) we have been able to explain all observed association in the HLA region with systemic sclerosis or its most frequent auto-antibodies: anti-topoisomerase I and anti-centromere auto-antibodies.

Introduction

SSc is a complex autoimmune disease characterized by vascular damage, altered immune responses and extensive fibrosis of skin and internal organs. The disease is clinically heterogeneous, and the two major and mutually exclusive subgroups are defined by presence of the auto-antibodies: DNA topoisomerase I (ATA), and anti-centromere auto-antibodies (CENP A and/or B proteins) (ACA) [1-3]. These autoantibodies correlate with the two clinical subgroups of SSc, namely limited cutaneous SSc (lcSSc, associated with ACA antibody), and diffuse cutaneous SSc (dcSSc, associated with ATA antibody). Around 90% of patients present one of these antibodies. Other less frequent auto-antibodies can also be found in these patients, *e.g.*, RNA polymerase III (pol-III), U3-RNP (fibrillarin), Th/To, PM/SCL, and U1-anti-ribonucleoprotein (RNP) [4-6]. In previous reports, several HLA class II alleles have been associated with either SSc overall or the auto-antibody subgroups [4, 7-10]. For example, HLA-DRB1*0101 and HLA-DQB1*0501 association is confined to ACA positive subgroup, whereas HLA-DPB1*1301 association is found in the ATA positive subgroup [4, 11, 12]. On the other hand HLA-DRB1*0701 has been associated with protection to SSc overall [4, 13]. Finally, recent GWAS studies of SSc show a strong association to the HLA class II region in the MHC. Unfortunately, the described associations of classical HLA alleles do not comprehensively explain the HLA associations detected in GWAS nor do they clarify which variants are associated with the disease. Here, we build upon available dense SNP data from our previous GWAS of SSc, using an imputation method designed for the MHC region to investigate the strong association of the HLA region with SSc. We aimed at defining the HLA alleles that explain

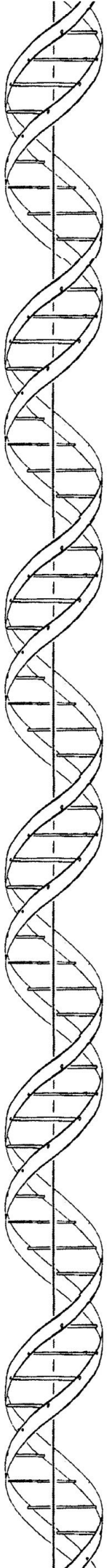


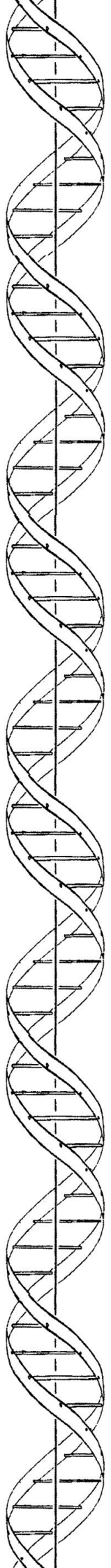
the association of the HLA region with SSc and to pinpoint the putative causal variations that drive these associations.



Results

Dense genotype data for the extended MHC region was obtained from our previous GWAS of SSC and used for imputation. The imputation method has been described elsewhere and has been used to map the HLA association in HIV, anti-CCP positive rheumatoid arthritis and ulcerative colitis [14-16]. In brief, the method uses a large data set reference panel of 2,767 individuals of European descent [17] with HLA class I and II molecules four digit typing and the genotypes of more than 7,500 common SNPs and deletion-insertion polymorphisms across the xMHC [18]. Imputation was performed with the program BEAGLE using standard setting (B. L. Browning, S. R. Browning, Am J Hum Genet 84, 210 (Feb, 2009)). Imputed HLA alleles or amino acids dosages were tested for association using a logistic regression model that corrects for population substructure and genotyping batch using PLINK version 1.07 [19]. For amino acid positions with >2 alleles, we used the omnibus test in the conditional haplotype analysis module in PLINK. We imputed the 3,841 SNPs, 894 polymorphic aminoacids and the *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DPA1*, *HLA-DPB1*, *HLA-DQA1*, *HLA-DQB1* and *HLA-DRB1* alleles of 2,296 SSc patients and 5,356 healthy controls. We checked the accuracy of the method in our dataset by comparing imputed HLA classical alleles with partially genotyped alleles in a smaller subset of 919 cases and controls, observing a 92% and an 83% of accuracy in the US and Spain cohorts, respectively (*supplementary table 1*). These accuracy rates are certainly too low for individual prediction of HLA type. However, inaccuracies will only reduce our power and do not cause spurious results, given that errors occur at random. Manual inspection showed no evidence for allele specific inaccuracies, suggesting random error.



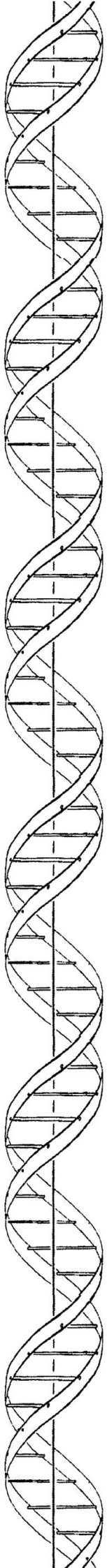


Association analysis of imputed data confirmed previous reported associations of *HLA-DRB1*0701*, *HLA-DPB1*1301* and *HLA-DRB1*1104* (table 1 and supplementary table 2). Remarkably, each association showed restriction to a single auto-antibody positive subsets of SSc patients (table 1). In overall SSc the most associated variants were SNP CHR6_POS32660045 and *HLA-DPB1*1301* at *HLA-DRB1* and *HLA-DPB1* respectively. These two peaks of association overlap with the location of the most significant variants observed in ACA and ATA positive subgroups, namely aminoacid position 13 and position 76 at *HLA-DRβ1* and *HLA-DPβ1* respectively. Furthermore, in the remaining subgroup of patients that are neither positive for ACA nor ATA the most associated variant was aminoacid position 67 of the *HLA-DRβ1* molecule (figure 1). In this subgroup, no significant association remained for the most associated variants of ACA and ATA positive subgroups.

We proceeded with stepwise conditional association analysis to pinpoint the alleles causing the observed associations in ACA and ATA subgroups (figure 1b-f and supplementary table 3). In ACA positive patients, the most associated variant was the aminoacid at position 13 in *HLA-DRβ1* (omnibus test $P = 6.15 \times 10^{-57}$). Subsequent conditioning steps revealed that association at aminoacids 60, and 71 significantly improved the goodness of fit of the logistic model; furthermore, the model including these three polymorphic aminoacid positions explained most of the association between the HLA region and ACA positive SSc (figure 1b-c and supplementary table 3). Conditional analysis of these aminoacids in different orders remained significant, indicating that indeed these three aminoacids are necessary to explain most of the association in this SSc subgroup.

For the ATA positive subgroup the most associated variant was an aminoacid at the position 76 of HLA-DP β 1 (*figure 1d-f* and *supplementary table 4*). Conditioning for this aminoacid explained all associations at HLA-DP β 1, and revealed subsequent independent associations at HLA-DR β 1 at aminoacids at position 58, 67, and 86. All positions improved the goodness of fit of the logistic model, and together explained most of the association in the MHC region in this auto-antibody subgroup (*figure 1d-f* and *supplementary table 4*). The residual MHC association was mainly carried by the classical HLA allele *HLA-DRB1*1104*. However, this imputed allele did not improve the goodness of fit of the model significantly (likelihood P value = 1.18×10^{-3}), nor was it in strong LD with any of the remaining variants. Again conditional testing of the aminoacids in different orders remained significant.

The seven aminoacids together correlate to the classical HLA alleles that were found associated to SSc previously (*table 3*). Furthermore, we performed a logistic regression analysis conditioned to the presence of the risk alleles of these seven aminoacids in the full SSc cohort, the non-ACA and ATA producing auto-antibodies (hence called double negatives), and both cutaneous subtypes of the disease (lcSSc and dcSSc). As observed in *supplementary table 5, figure 1* and *supplementary figure 1* all association in SSc or any of its subgroups in the MHC was fully conditioned by the seven aminoacids.



Discussion

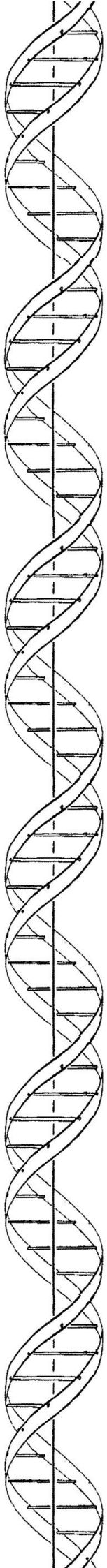
Here, we identified seven variable aminoacids that collectively explain the majority of the HLA association with SSc and correlate to previous described classical HLA allele associations. The one exception of the latter is the *HLA-DRB1*1602* allele previously described to be associated in the Choctaw Native American tribe, which could not be imputed in our cohorts due to its low frequency in European ancestry populations [20]. Furthermore, Karp *et al.* found aminoacids 26, 28, 30, 37, 67, 70, 71 and 86 to be the causative variation for the observed association of *HLA-DRB1*1104* with SSc using a different approach to fine map the HLA-DR β 1 association. However, we were able to impute all of the aminoacidic positions described by Karp *et al.*, but none of these positions showed an independent association with SSc after conditional analysis with above identified loci. Only aminoacid 86 was also identified in our data as part of the causal variations (*tables 2 and 3, supplementary tables 2, 3 and 4*).

Our results suggest that the aminoacids identified in autoantibody producing subgroups drive the majority, if not all, of the MHC association with SSc. The strong association of the two groups of aminoacids according to autoantibody status also suggests that the observed HLA association in the double negative group may have been caused by subjects in which autoantibodies escaped detection. Nevertheless, we limited power to rule out the possibility of other aminoacids explaining the association for other, less frequent, autoantibodies found in SSc patients, such as anti-RNA polymerase III autoantibodies.

All identified associated aminoacid positions are located in peptide binding sites of either the HLA-DR β 1 or HLA-DP β 1 molecules, and thus are probably affecting the affinity of the

HLA for the peptide [21, 22]. In the case of ACA positive patients these aminoacids (13, 60 and 71 of HLA-DR β 1) could influence the union of centromere components, *e.g.* CENP-A and CENP-B as previously described [23, 24]. On the other hand the corresponding aminoacid positions associated with ATA (position 76 of HLA-DP β 1 and 58, 67 and 86 of HLA-DR β 1) should be influencing the recognition of topoisomerase I epitopes. Furthermore, most identified aminoacid alleles associated with risk or protection had differential features regarding the side chains (*supplementary table 6*). It is also noteworthy that no independent association was found for any HLA class I with SSc or any of its considered subgroups.

In conclusion, we have been able to explain most of the observed association in the HLA region with SSc by the presence of specific alleles in the aminoacid positions 13, 58, 60, 67, 71 and 86 of the HLA-DR β 1 molecule and position 76 of the HLA-DP β 1 molecule, which influence the SSc risk through the auto-antibody (ACA and ATA) production risk. We add more evidence that SSc is not a genetically homogeneous disease. The two models of aminoacids define the ACA positive and ATA positive auto-antibody subgroups of the disease, account for the majority of HLA association in the MHC region, and explain the HLA alleles classically associated with the disease among patients of European ancestry.



Materials and Methods

GWAS data

The cohorts analyzed were composed of 2,296 Caucasian SSc patients and 5,356 healthy controls from USA, Spain, Germany and The Netherlands previously described, whose key features can be found in *supplementary table 7* [25]. All cases met the American College of Rheumatology preliminary criteria for the classification of SSc [26]. Furthermore, patients were classified according to the extent of skin involvement into limited (lcSSc) or diffuse (dcSSc) forms [27]. In addition, the presence of SSc specific auto-antibodies, anti-topoisomerase I (ATA, anti-Scl70) and anti-centromere (ACA) was assessed by passive immunodiffusion against calf thymus extract and indirect immunofluorescence of Hep-2 cells, respectively.

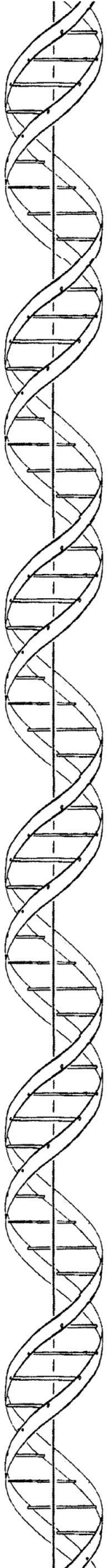
Collection of blood samples and clinical information from case and control subjects was undertaken with informed consent and relevant ethical review board approval from each contributing centre in accordance with the tenets of the Declaration of Helsinki.

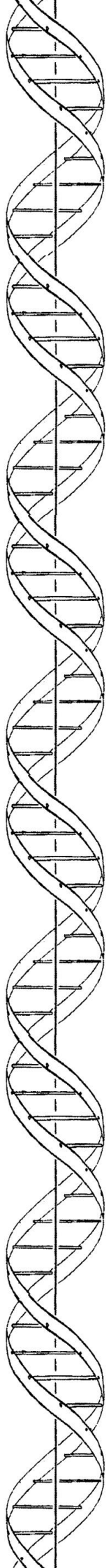
SNP data in the different cohorts was originally obtained from Illumina 550k and 370k HumanHap Illumina chips as previously described [25]. Data in the SSc GWAS cohorts prior to imputation was filtered as follows: Using Plink, we identified and excluded pairs of genetically related subjects or duplicates and excluded the genetic-pair members with lower call rates. To identify individuals who might have non-western European ancestry, we merged our case and control data with the data from the HapMap Project (60 western European (CEU), 60 Nigerian (YRI), 90 Japanese (JPT) and 90 Han Chinese (CHB) samples). We used principal component analysis as implemented in HelixTree, plotting the

first two principal components for each individual. All individuals who did not cluster with the main CEU cluster (defined as deviating more than 4 standard deviations from the cluster centroids) were excluded from subsequent analyses. Additional filters were applied as previously described [25]. Then we filtered for genotyping quality, removing SNPs with a genotyping success call rate $< 99\%$, individuals with a SNP success call rate $< 99\%$ and those showing $MAF < 1\%$. Deviation of the genotype frequencies in the controls from those expected under Hardy-Weinberg equilibrium was assessed by a χ^2 test or Fisher's exact test when an expected cell count was < 5 . SNPs deviating from Hardy-Weinberg equilibrium ($P < 0.001$) were eliminated from the study. The genotyping success call rate on the merged dataset after all these quality filters were applied was 99.83%. Once all filters were applied, we extracted for all individuals all SNP genotypes between the positions 20,000,000 and 40,000,000 in chromosome 6 (*i.e.* the xMHC) for imputation, resulting in a total of 2,892 SNPs after quality control.

Imputation

The 2,892 SNPs obtained in the xMHC after all quality controls were used for the imputation process. We used as reference panel for the imputation 2,767 individuals of European descent [17] with HLA class I and II molecules four digit typing and the genotypes of more than 7,500 common SNPs and deletion-insertion polymorphisms across the xMHC [18]. The imputation process was performed using the Beagle software [28]. Imputed data in the xMHC either for SNPs, aminoacids or the HLA four digits code were filtered as follows: variants with a success call rate below 95% were excluded, variants with a MAF below 1% were excluded and all individuals with a SNP success call rate below 95% were excluded. After these filters a total of 3,841 SNPs remained. Also, through



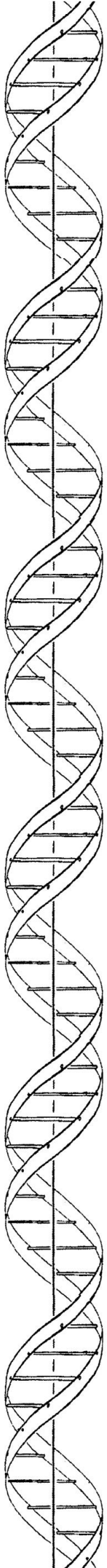


the imputation process a total of 894 polymorphic aminoacidic positions were obtained (143 for *HLA-A*, 197 for *HLA-B*, 118 for *HLA-C*, 34 for *HLA-DPB1*, 10 for *HLA-DPA1*, 92 for *HLA-DQB1*, 56 for *HLA-DQA1* and 244 for *HLA-DRB1*). At last, the alleles of the class I (*HLA-A*, *HLA-B* and *HLA-C*) and class II (*HLA-DPA1*, *HLA-DPB1*, *HLA-DQA1*, *HLA-DQB1* and *HLA-DRB1*) were obtained. The alleles of each of the class I and II molecules imputed can be found in *supplementary table 1*. We used previously genotyped HLA alleles partial data from 493 individuals from the US and 426 individuals from Spain for the *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DPA1*, *HLA-DPB1*, *HLA-DQA1*, *HLA-DQB1* and *HLA-DRB1* genes at 4 digits, which were partially included in previous reports [4, 10] in order to assess the accuracy of the imputation. This resulted in an accuracy of 92% in the US cohort and 83% in the Spanish cohort in the 4 digits comparison (*supplementary table 1*). In the US population the mean accuracy for all the HLA genes reached a higher value probably due to the fact that the reference panel used for imputation was also from US origin. Spanish cohort was more separated from the US (and also Netherlands and Germany) in the PC analysis as seen in *supplementary figure 2*, which could explain the slightly lower imputation accuracy in this population.

Statistical analysis

We performed the association analyses by the means of unconditioned and conditioned logistic regression analysis to account for dependency among the xMHC associations found. To control for population differences we included in the logistic regression models the country of precedence of the individuals as a covariate. To identify the independent signals, first, we performed a conditional logistic regression analysis using as condition the top-most associated variants in the unconditioned analysis, second, we used the top-most

associated variant independent of the first one as condition, third, we used the top-most associated variant independent from the first and the second ones as condition, and so on until we identified all independent variants in the xMHC region for SSc or any of its sub-phenotypes. Since the present work is based on GWAS data, P values $< 5 \times 10^{-8}$ were considered significant in order to minimize the type I errors. Odds ratios (OR) were calculated according to Woolf's method. In this stepwise conditional logistic regression analysis we searched recursively for models which better explained all association present in the xMHC. When more than one model was found to explain the association of one HLA allele, we compared the goodness of fit of each one. We assessed the significance of the improvement in fit by calculating the deviance (defined as $-2 \times$ the log likelihood), which follows a χ^2 distribution. We assumed that the model was not improved when the improvement of fit was not significantly higher than the previous model. The multi-allelic variants analyzed in the present study were encoded as binary variables according to the presence or absence of each of the alleles for each aminoacidic position or HLA molecule. For example, in the case HLA-A 4 digits alleles, the allele HLA-A*0101 was binary encoded as the presence or absence of HLA-A*0101, the allele HLA-A*0201 was binary encoded as the presence or absence of HLA-A*0201 and so on. These variables were analyzed by including in one single logistic regression model all possible alleles at each position, which gave an omnibus significance value for the locus. All analyses were performed using Plink software (Version 1.07) [19] (<http://pngu.mgh.harvard.edu/~purcell/plink/>) and HelixTree SNP Variation Suite (Version 7) (<http://www.goldenhelix.com/>). The imputation process was performed using Beagle software [28]. The 3D models of the HLA molecules were done with UCSF Chimera [29].

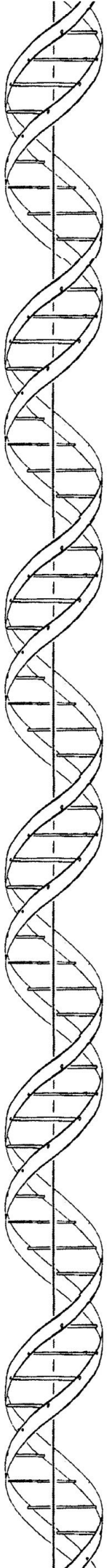


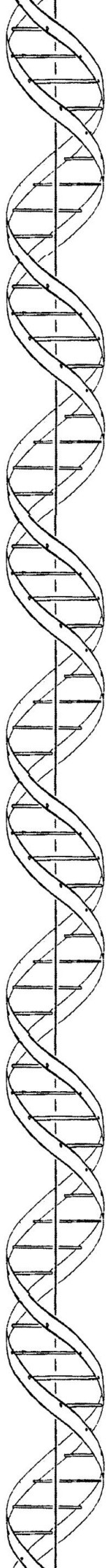
Acknowledgements and Funding

We thank Sofia Vargas, Sonia Garcia and Gema Robledo for their excellent technical assistance, and all the patients and healthy controls for kindly accepting their essential collaboration. We would also like to thank the following organizations: the EULAR Scleroderma Trials and Research (EUSTAR), the German Network of Systemic Sclerosis and Banco Nacional de ADN (University of Salamanca, Spain). This work was supported by the following grants: **J.M.** was funded by GEN-FER from the Spanish Society of Rheumatology, SAF2009-11110 from the Spanish Ministry of Science, CTS-4977 from Junta de Andalucía (Spain), Redes Temáticas de Investigación Cooperativa Sanitaria Program, RD08/0075 (RIER) from Instituto de Salud Carlos III (ISCIII, Spain) and Fondo Europeo de Desarrollo Regional (FEDER). **T.R.D.J.R.** was funded by the VIDI laureate from the Dutch Association of Research (NWO) and Dutch Arthritis Foundation (National Reumafonds). **J.M.** and **T.R.D.J.R.** were sponsored by the Orphan Disease Program grant from the European League Against Rheumatism (EULAR). **B.P.C.K.** is supported by the Dutch Diabetes Research Foundation (grant 2008.40.001) and the Dutch Arthritis Foundation (Reumafonds, grant NR 09-1-408). **F.K.T.**, **F.C.A.**, **S.A.** and **M.D.M.** were supported by NIH/NIAMS Scleroderma Family Registry and DNA Repository (N01-AR-0-2251), NIH/NIAMS-RO1- AR055258, and NIH/NIAMS Center of Research Translation in Scleroderma (1P50AR054144), the Department of Defense Congressionally Directed Medical Research Programs (W81XWH-07-01-0111), and K23-AR-061436. **J.M.** and **X.Z.** were supported by the grant NIH NIAID 1U01AI09090-01.

References

1. Gabrielli, A., E.V. Avvedimento, and T. Krieg, *Scleroderma*. N Engl J Med, 2009. **360**(19): p. 1989-2003.
2. Martin, J.E., L. Bossini-Castillo, and J. Martin, *Unraveling the genetic component of systemic sclerosis*. Hum Genet, 2012.
3. Steen, V.D., *The many faces of scleroderma*. Rheum Dis Clin North Am, 2008. **34**(1): p. 1-15; v.
4. Arnett, F.C., et al., *Major histocompatibility complex (MHC) class II alleles, haplotypes and epitopes which confer susceptibility or protection in systemic sclerosis: analyses in 1300 Caucasian, African-American and Hispanic cases and 1000 controls*. Ann Rheum Dis, 2010. **69**(5): p. 822-7.
5. Sharif, R., et al., *Anti-fibrillarin antibody in African American patients with systemic sclerosis: immunogenetics, clinical features, and survival analysis*. J Rheumatol. **38**(8): p. 1622-30.
6. Steen, V., et al., *A clinical and serologic comparison of African-American and Caucasian patients with systemic sclerosis*. Arthritis Rheum.
7. Genth, E., et al., *Immunogenetic associations of scleroderma-related antinuclear antibodies*. Arthritis Rheum, 1990. **33**(5): p. 657-65.
8. Gladman, D.D., et al., *Increased frequency of HLA-DR5 in scleroderma*. Arthritis Rheum, 1981. **24**(6): p. 854-6.
9. Gorlova, O., et al., *Identification of novel genetic markers associated with clinical phenotypes of systemic sclerosis through a genome-wide association strategy*. PLoS Genet, 2011. **7**(7): p. e1002178.
10. Karp, D.R., et al., *Novel sequence feature variant type analysis of the HLA genetic association in systemic sclerosis*. Hum Mol Genet, 2010. **19**(4): p. 707-19.
11. Reveille, J.D., et al., *Association of amino acid sequences in the HLA-DQB1 first domain with antitopoisomerase I autoantibody response in scleroderma (progressive systemic sclerosis)*. J Clin Invest, 1992. **90**(3): p. 973-80.
12. Reveille, J.D., et al., *Association of polar amino acids at position 26 of the HLA-DQB1 first domain with the anticentromere autoantibody response in systemic sclerosis (scleroderma)*. J Clin Invest, 1992. **89**(4): p. 1208-13.
13. Simeon, C.P., et al., *Association of HLA class II genes with systemic sclerosis in Spanish patients*. J Rheumatol, 2009. **36**(12): p. 2733-6.
14. Raychaudhuri, S., et al., *Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis*. Nat Genet, 2012. **44**(3): p. 291-6.
15. Achkar, J.P., et al., *Amino acid position 11 of HLA-DRbeta1 is a major determinant of chromosome 6p association with ulcerative colitis*. Genes Immun, 2012. **13**(3): p. 245-52.
16. Pereyra, F., et al., *The major genetic determinants of HIV-1 control affect HLA class I peptide presentation*. Science, 2010. **330**(6010): p. 1551-7.
17. Brown, W.M., et al., *Overview of the MHC fine mapping data*. Diabetes Obes Metab, 2009. **11 Suppl 1**: p. 2-7.
18. de Bakker, P.I., et al., *A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC*. Nat Genet, 2006. **38**(10): p. 1166-72.
19. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses*. Am J Hum Genet, 2007. **81**(3): p. 559-75.



- 
20. Arnett, F.C., et al., *Autoantibodies to fibrillarin in systemic sclerosis (scleroderma). An immunogenetic, serologic, and clinical analysis.* Arthritis Rheum, 1996. **39**(7): p. 1151-60.
 21. Brown, J.H., et al., *Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1.* Nature, 1993. **364**(6432): p. 33-9.
 22. Dai, S., et al., *Crystal structure of HLA-DP2 and implications for chronic beryllium disease.* Proc Natl Acad Sci U S A. **107**(16): p. 7425-30.
 23. Kremer, L., et al., *Proteins responsible for anticentromere activity found in the sera of patients with CREST-associated Raynaud's phenomenon.* Clin Exp Immunol, 1988. **72**(3): p. 465-9.
 24. Mahler, M., et al., *Development of a CENP-A/CENP-B-specific immune response in a patient with systemic sclerosis.* Arthritis Rheum, 2002. **46**(7): p. 1866-72.
 25. Radstake, T.R., et al., *Genome-wide association study of systemic sclerosis identifies CD247 as a new susceptibility locus.* Nat Genet, 2010. **42**(5): p. 426-9.
 26. *Preliminary criteria for the classification of systemic sclerosis (scleroderma). Subcommittee for scleroderma criteria of the American Rheumatism Association Diagnostic and Therapeutic Criteria Committee.* Arthritis Rheum, 1980. **23**(5): p. 581-90.
 27. LeRoy, E.C. and T.A. Medsger, Jr., *Criteria for the classification of early systemic sclerosis.* J Rheumatol, 2001. **28**(7): p. 1573-6.
 28. Browning, S.R. and B.L. Browning, *Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering.* Am J Hum Genet, 2007. **81**(5): p. 1084-97.
 29. Pettersen, E.F., et al., *UCSF Chimera--a visualization system for exploratory research and analysis.* J Comput Chem, 2004. **25**(13): p. 1605-12.

Legends to figures

Figure 1. Manhattan plots of the xMHC region representing the $-\text{Log}_{10}$ of the P values for the total SSc as well as the auto-antibody positive subgroups. a) unconditioned joined Manhattan plot, b) ACA+ unconditioned Manhattan plot, c) ACA+ Manhattan plot controlling for the association found in *HLA-DRB1*, d) ATA+ unconditioned Manhattan plot, e) ATA+ Manhattan plot controlling for the association found in *HLA-DPB1*, f) ATA+ Manhattan plot controlling for both the associations found in *HLA-DPB1* and *HLA-DRB1*, g) double negative unconditioned Manhattan plot, and h) double negative Manhattan plot conditioned for HLA-DRB1's association in ACA+ and HLA-DRB1/DPB1's association in ATA+.

Figure 2. 3D models of each of the molecules causing the genetic predisposition to develop auto-antibodies in systemic sclerosis patients. The aminoacidic positions which best explain all observed associations in xMHC are highlighted in turquoise blue. a) Aminoacidic positions responsible for the association of HLA-DRB1 with the ACA+ subgroup, b) aminoacidic positions responsible for the association of HLA-DPB1 with the ATA+ subgroup, c) aminoacidic positions responsible for the association of HLA-DRB1 with the ATA+ subgroup.

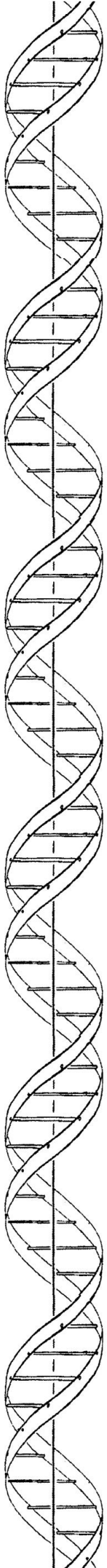


Table 1. HLA classic four digits alleles independently associated with systemic sclerosis, more specifically, with its auto-antibody positive subgroups. Significant *P* values are marked in bold.

HLA Gene	Allele	<i>P</i> SSc	OR SSc	<i>P</i> ACA+	OR ACA+	<i>P</i> ACA-	OR ACA-	<i>P</i> ATA+	OR ATA+	<i>P</i> ATA-	OR ATA-	Association Group
<i>HLA-DRB1</i>	0701	8.77x10⁻¹⁶	0.61	7.73x10⁻²⁰	0.31	8.71x10 ⁻⁵	0.77	4.41x10 ⁻¹	0.93	7.87x10⁻¹⁹	0.54	ACA+
<i>HLA-DRB1</i>	0801	2.69x10⁻⁹	1.78	1.39x10⁻²³	3.28	4.65x10 ⁻¹	1.10	2.81x10 ⁻¹	1.25	1.64x10⁻¹⁰	1.91	ACA+
<i>HLA-DQB1</i>	0302	4.97x10 ⁻⁵	1.26	1.67x10⁻¹¹	1.71	4.57x10 ⁻¹	1.05	2.98x10 ⁻²	0.75	3.12x10⁻⁸	1.39	ACA+
<i>HLA-DQB1</i>	0501	4.66x10 ⁻⁴	1.20	1.02x10⁻²⁵	2.10	5.60x10 ⁻³	0.83	1.79x10 ⁻⁷	0.48	6.44x10⁻¹⁰	1.39	ACA+
<i>HLA-DPB1</i>	1301	2.03x10⁻²²	2.73	6.94x10 ⁻¹	1.08	4.78x10⁻³³	3.65	6.97x10⁻⁷⁷	10.80	6.55x10 ⁻²	1.28	ATA+
<i>HLA-DRB1</i>	1104	3.51x10⁻¹³	2.12	2.95x10 ⁻²	1.45	7.64x10⁻¹⁶	2.47	3.22x10⁻³⁴	5.47	5.88x10 ⁻³	1.40	ATA+

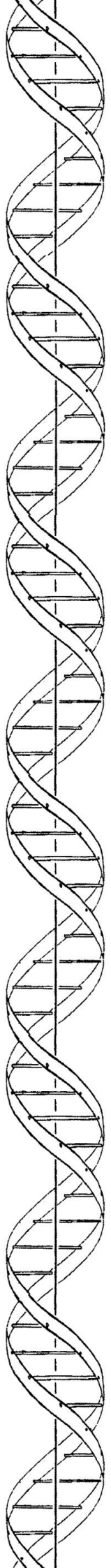


Table 2. Aminoacids independently associated with SSc which explain all observed association of the HLA classic alleles with the auto-antibody positive subgroups. †we only show here the associations which proved to be independent in the present study, for any other previously described HLA association see *table 3*, and for the specific conditioned *P* values see *supplementary tables 2 and 3*.

SSc subgroup	HLA gene	Position	Allele	Aminoacid		<i>P</i> value	OR	CI 95%	Conditions the association of
				Freq. Cases	Freq. Controls				
ACA+	<i>HLA-DRB1</i>	13	HFG	0.528	0.335	2.89x10 ⁻⁴⁵	2.21	1.98-2.47	HLA-DRB1*0701
		60	S	0.066	0.156	1.37x10 ⁻¹⁹	0.38	0.31-0.47	HLA-DRB1*0801
		71	R	0.407	0.495	2.02x10 ⁻¹⁰	0.70	0.63-0.78	HLA-DQB1*0302 HLA-DQB1*0501
ATA+	<i>HLA-DPB1</i>	76	I	0.150	0.023	3.61x10 ⁻⁷⁰	8.69	6.84-11.04	HLA-DPB1*1301
ATA+	<i>HLA-DRB1</i>	58	E	0.249	0.111	5.60x10 ⁻³²	2.82	2.37-3.35	
		86	V	0.539	0.472	1.02x10 ⁻⁴	1.32	1.15-1.51	HLA-DRB1*1104
		67	L	0.224	0.406	7.35x10 ⁻²⁵	0.43	0.36-0.50	

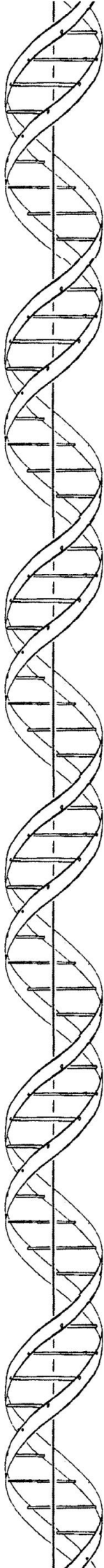


Table 3. All MHC class II alleles previously reported to be associated with SSc or any of its subgroups analyzed in the present study. †The best *P* value from all subsets considered, *i.e.* SSc, ACA and ATA. ‡The conditioning of each HLA allele was only performed when the unconditioned *P* value was significant. *HLA alleles previously reported as SSc genetic risk factors have been narrowed down to auto-antibody positive subgroups and fully explained within them in the present study (see *table 1*).

Gene	Variation	Subgroup	Ethnicity	Best <i>P</i> value	OR (CI 95%)	Present Study		
						Subgroup	Conditioned <i>P</i> value†	Conditioning Variation
<i>HLA-DRB1</i>	01(01)	ACA	Hispanic, Japanese and Caucasians	5.40x10⁻¹⁹	2.01 (1.72-2.34)	ACA	9.15x10 ⁻²	AAs 43, 90 and 101 of HLA-DRB1
<i>HLA-DRB1</i>	04(02)	ACA	Hispanic, Japanese and Caucasians	1.86x10 ⁻¹	1.29 (0.88-1.88)	ACA	NA	NA
<i>HLA-DQB1</i>	05(01)	ACA	Hispanic, Japanese and Caucasians	1.02x10⁻²⁵	2.10 (1.83-2.41)	ACA	1.60x10 ⁻²	AAs 43, 90 and 101 of HLA-DRB1
<i>HLA-DRB1</i>	1101	ATA	Caucasians and African Americans	3.79x10 ⁻⁷	1.73 (1.40-2.14)	ATA	NA	NA
<i>HLA-DRB1</i>	11(04)	ATA	Japanese and Caucasians	3.22x10⁻³⁴	5.48 (4.16-7.18)	ATA	4.82x10 ⁻⁶	AAs 88, 97 and 116 of HLA-DRB1
<i>HLA-DRB1</i>	1502	ATA	Japanese	1.37x10 ⁻²	2.01 (1.15-3.50)	ATA	NA	NA
<i>HLA-DQB1</i>	0301	ATA	Caucasian and Black	5.12x10⁻¹³	1.78 (1.52-2.07)	ATA	3.62x10 ⁻¹	AAs 88, 97 and 116 of HLA-DRB1
<i>HLA-DQB1</i>	0601	ATA	Caucasian, Japanese and Black	1.24x10 ⁻²	2.03 (1.17-3.54)	ATA	NA	NA
<i>HLA-DPB1</i>	1301	ATA	Caucasians	6.97x10⁻⁷⁷	10.80 (8.40-13.89)	ATA	2.09x10 ⁻⁵	AA 105 of HLA-DPB1
<i>HLA-DQB1</i>	05(01)	ATA	Caucasians	1.79x10 ⁻⁷	0.48 (0.37-0.63)	ATA	NA	NA
<i>HLA-DPB1</i>	0901	ATA	Japanese	2.89x10 ⁻¹	1.40 (0.75-2.62)	ATA	NA	NA
<i>HLA-DRB1</i>	1104	SSc	Caucasians	3.22x10⁻³⁴	5.48 (4.16-7.18)	ATA*	4.82x10 ⁻⁶	AAs 88, 97 and 116 of HLA-DRB1
<i>HLA-DRB1</i>	0701	SSc	Caucasians	7.73x10⁻²⁰	0.31 (0.24-0.40)	ACA*	6.71x10 ⁻¹	AAs 43, 90 and 101 of HLA-DRB1
<i>HLA-DRB1</i>	1501	SSc	Caucasians	3.24x10 ⁻⁵	0.68 (0.57-0.82)	ACA*	NA	NA

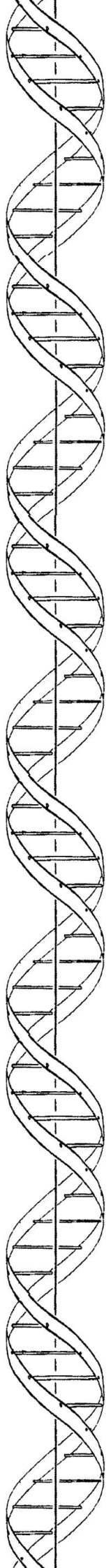


Figure 1.

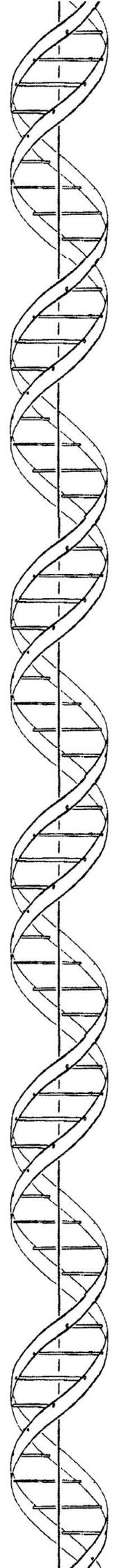
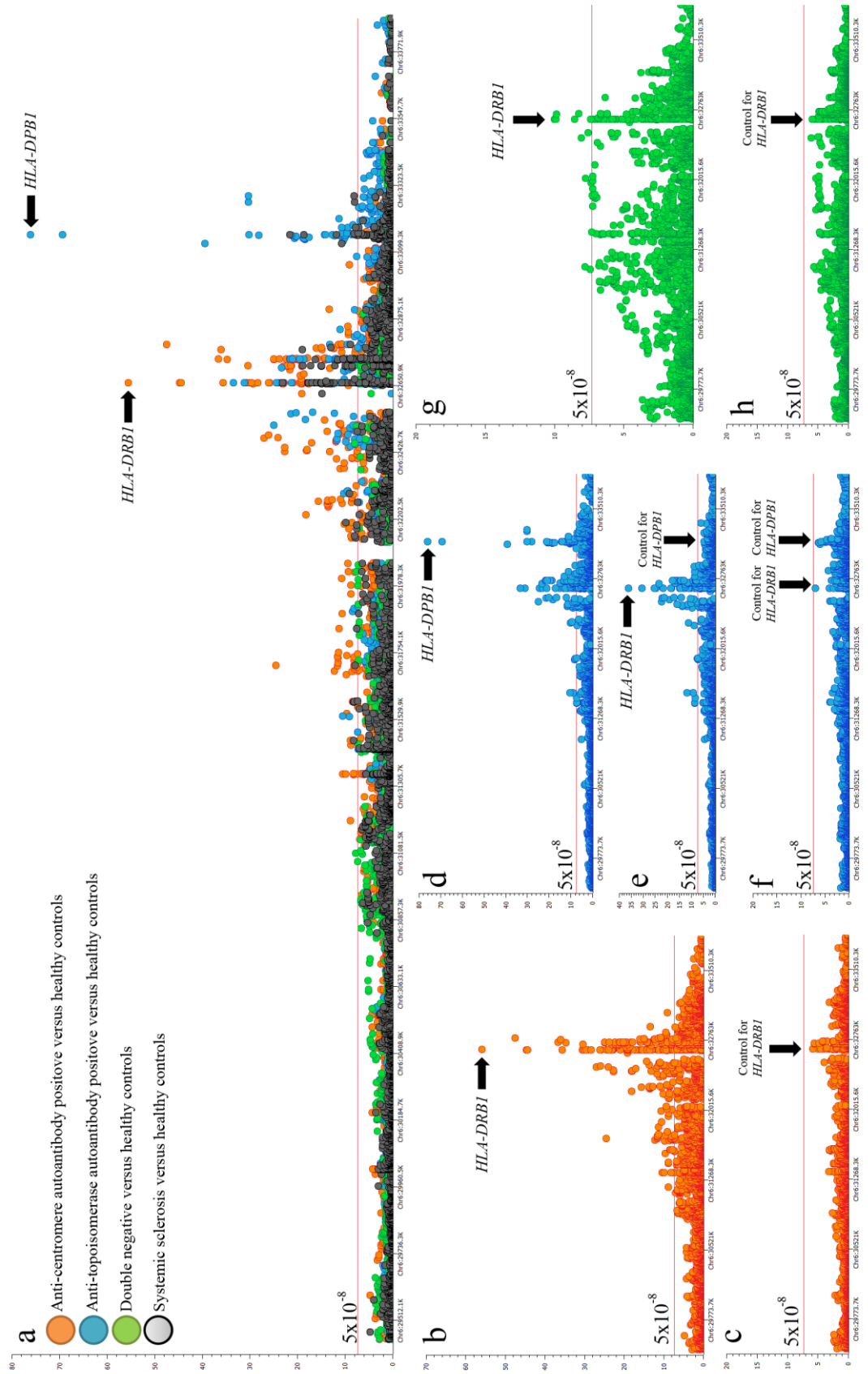
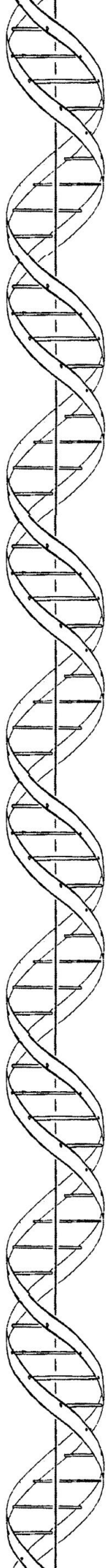
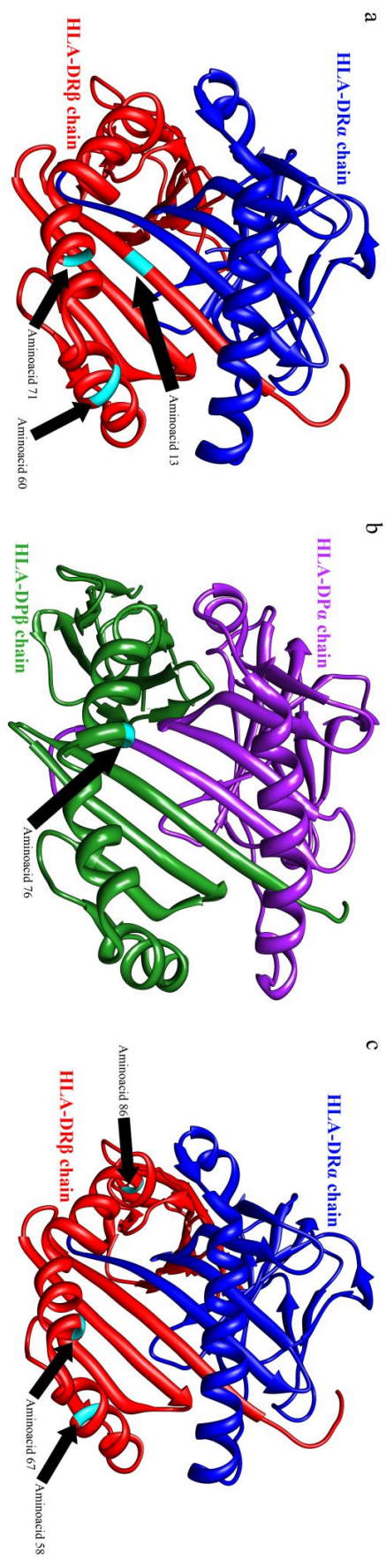


Figure 2.



Legend to supplementary tables and figures

Supplementary table 1. Alleles imputed for each HLA gene included in the present study and imputation accuracy of each gene separately.

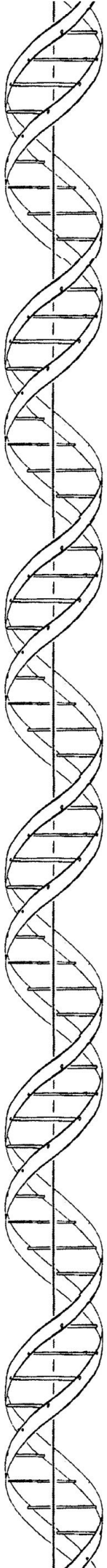
Supplementary table 2. Detailed analysis of all variations (SNPs, aminoacids and HLA alleles) analyzed in the overall SSc and the ACA+, ATA+, double negatives, lcSSc and dcSSc subgroups. The HLA-DRB1 aminoacids indentified as causative by Karp et al. are marked in **green**. The nomenclature of the variations is explained in *supplementary note 1*.

Supplementary table 3. Detailed analysis of all variations (SNPs, aminoacids and HLA alleles) analyzed in the ACA positive subset of patients. The nomenclature of the variations is explained in *supplementary note 1*.

Supplementary table 4. Detailed analysis of all variations (SNPs, aminoacids and HLA alleles) analyzed in the in the ATA positive subset of patients. The nomenclature of the variations is explained in *supplementary note 1*.

Supplementary table 5. Detailed analysis of all variations (SNPs, aminoacids and HLA alleles) conditioned to our seven aminoacid model in SSc and the ACA+, ATA+, double negatives, lcSSc and dcSSc subgroups. The nomenclature of the variations is explained in *supplementary note 1*.

Supplementary table 6. Key features of the different alleles imputed in our populations of the HLA class II molecules aminoacidic positions which explain all association of the HLA region with SSc or any of its considered subgroups. *Frequency for our control



populations. **Arbitrary feature of the considered alleles which differences risk/protection/neutral alleles at each position.

Supplementary table 7. Key features of the cohorts used in this study.

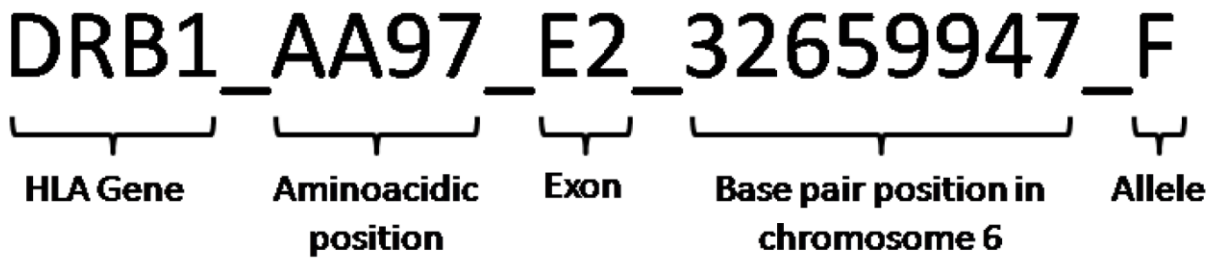
Supplementary figure 1. Manhattan plots of the xMHC region representing the $-\text{Log}_{10}$ of the P values for the total SSc as well as the cutaneous subtypes of the disease. a) unconditioned joint Manhattan plot, b) total SSc unconditioned Manhattan plot, c) SSc Manhattan plot controlling for the seven aminoacid model, d) lcSSc unconditioned Manhattan plot, e) lcSSc Manhattan plot controlling for the seven aminoacid model, f) dcSSc unconditioned Manhattan plot, and g) dcSSc Manhattan plot controlling for the seven aminoacid model.

Supplementary figure 2. Principal component plot for the first two eigenvectors. Every represented individual who deviated more than four standard deviations from their populations centroid were excluded from further analysis as principal component outliers.

Supplementary note 1

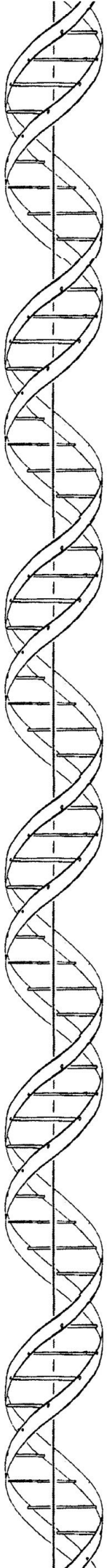
SNPs are named after their rs# number or, if not available, after their position in chromosome six build 36 (e.g. CHR6_POS32660045).

Aminoacid variations are named after the HLA gene in which they are found, aminoacidic position within the protein, exon in which the aminoacid is located, base position on chromosome six and considered alleles in one letter aminoacidic code. When more than one allele is in the nomenclature, it refers to the presence of any of them.

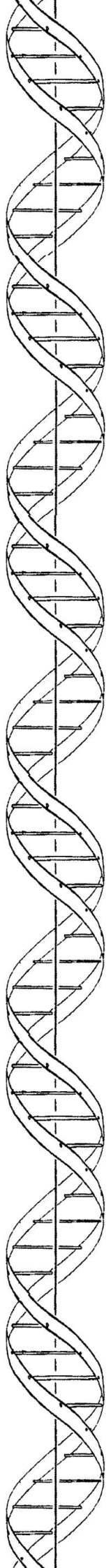
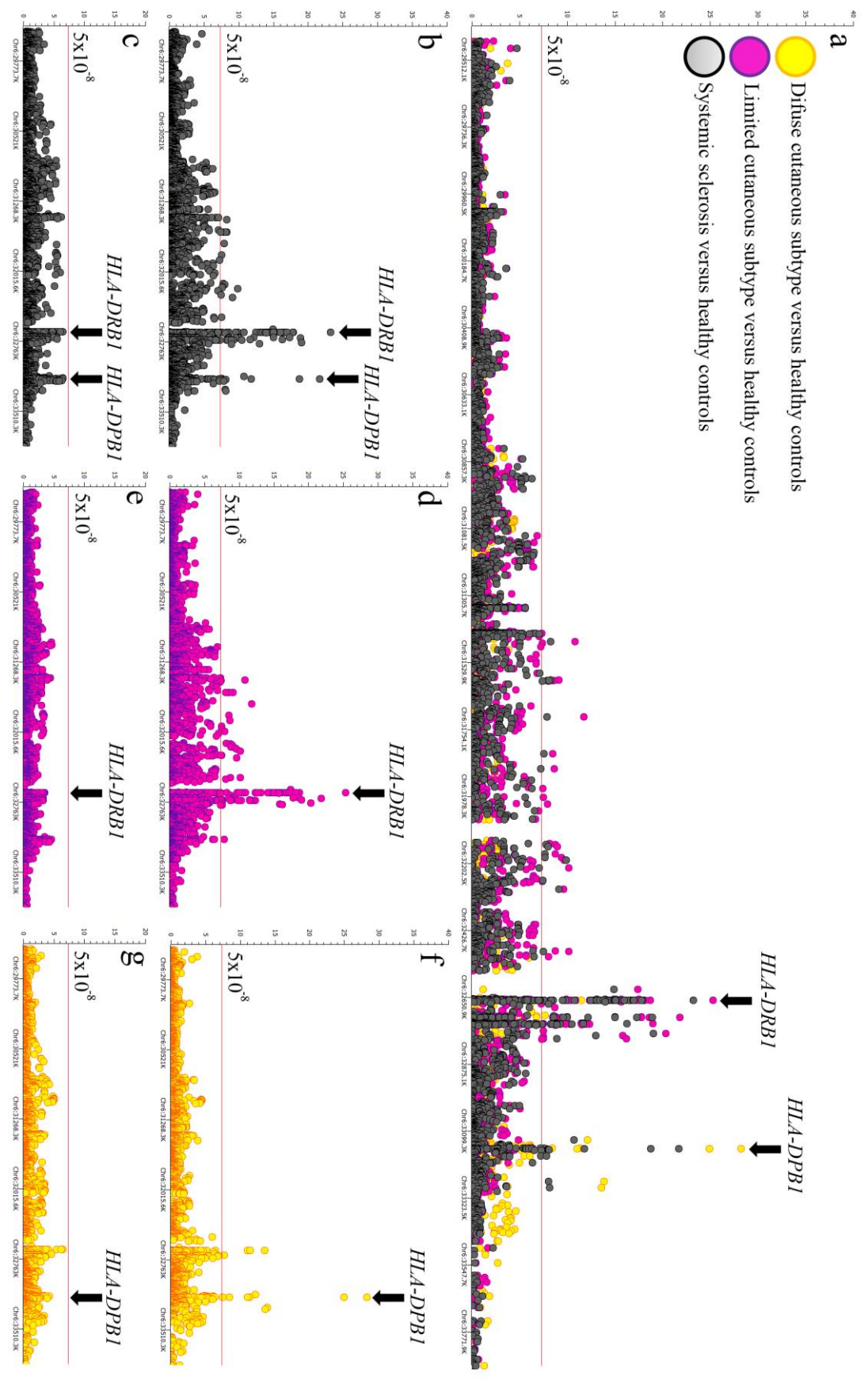


For the imputed amino acids, there exist many in the leading sequence of the HLA protein that don't make it into the final protein. Thus systematic changes must be made to the numbering system: in HLA-A subtract 24, in HLA-B subtract 25, HLA-C subtract 25, in HLA-DRB1 subtract 30, in HLA-DPA1 subtract 30, in HLA-DPB1 subtract 29, in HLA-DQA1 subtract 23 and in HLA-DQB1 subtract 33. Hence, in the example above (DRB1_AA97_E2_32659947_F) the position of the aminoacid in the protein is 67.

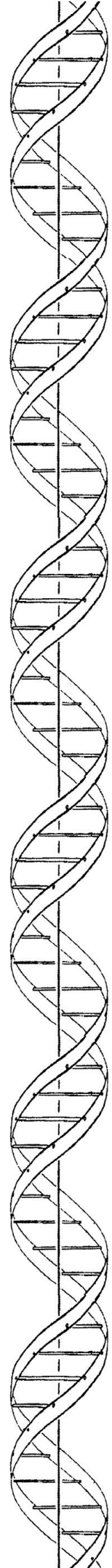
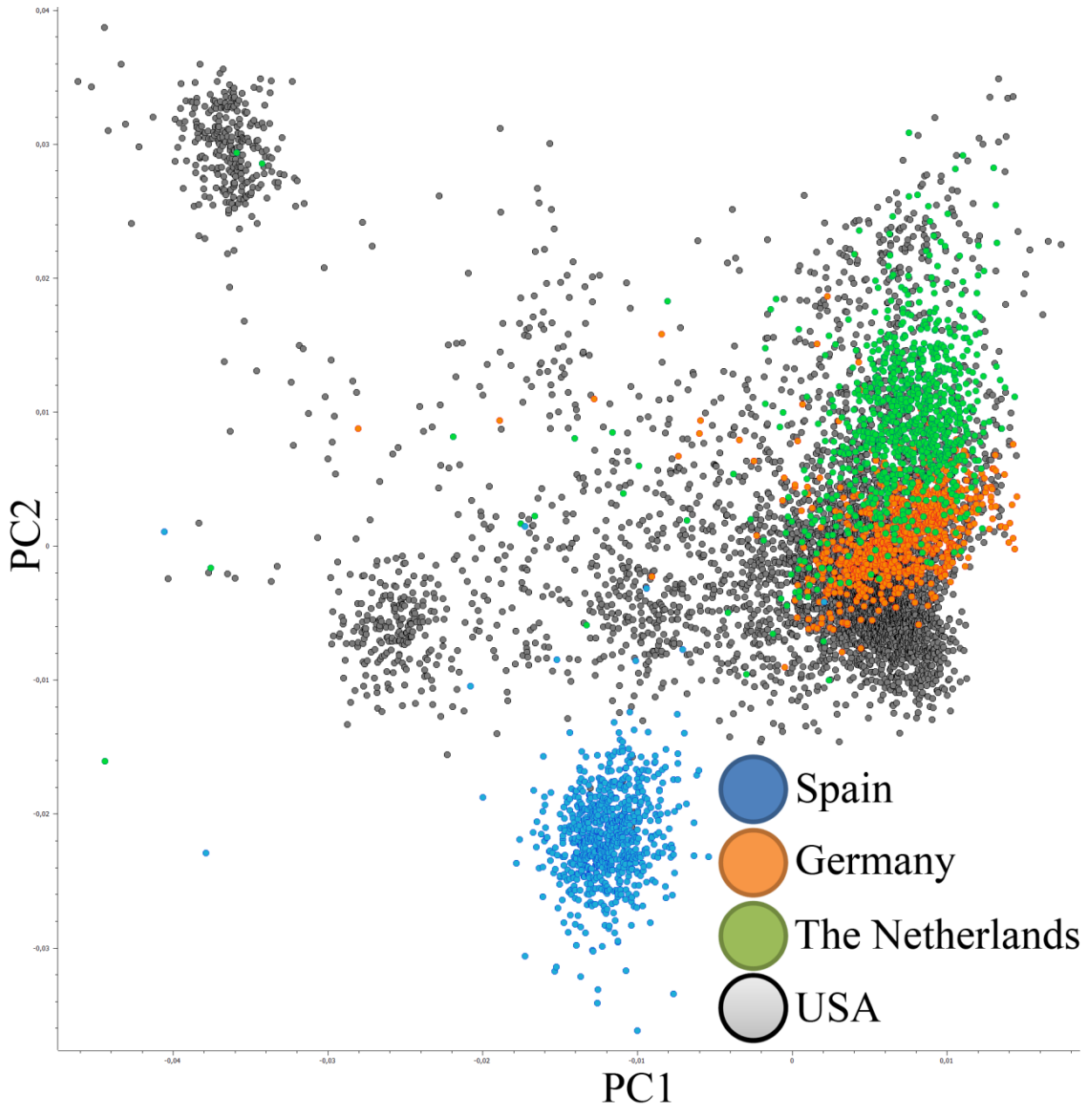
HLA alleles are named after the HLA gene and the considered allele (e.g. HLA_DPB1*1301).

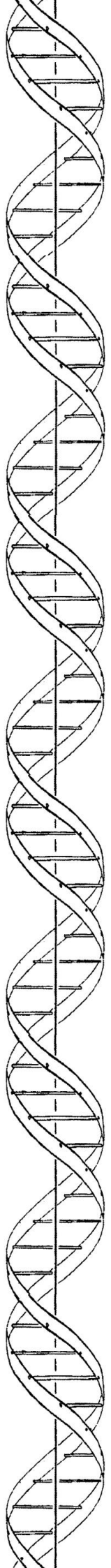


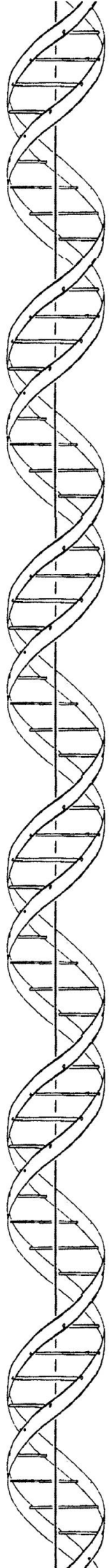
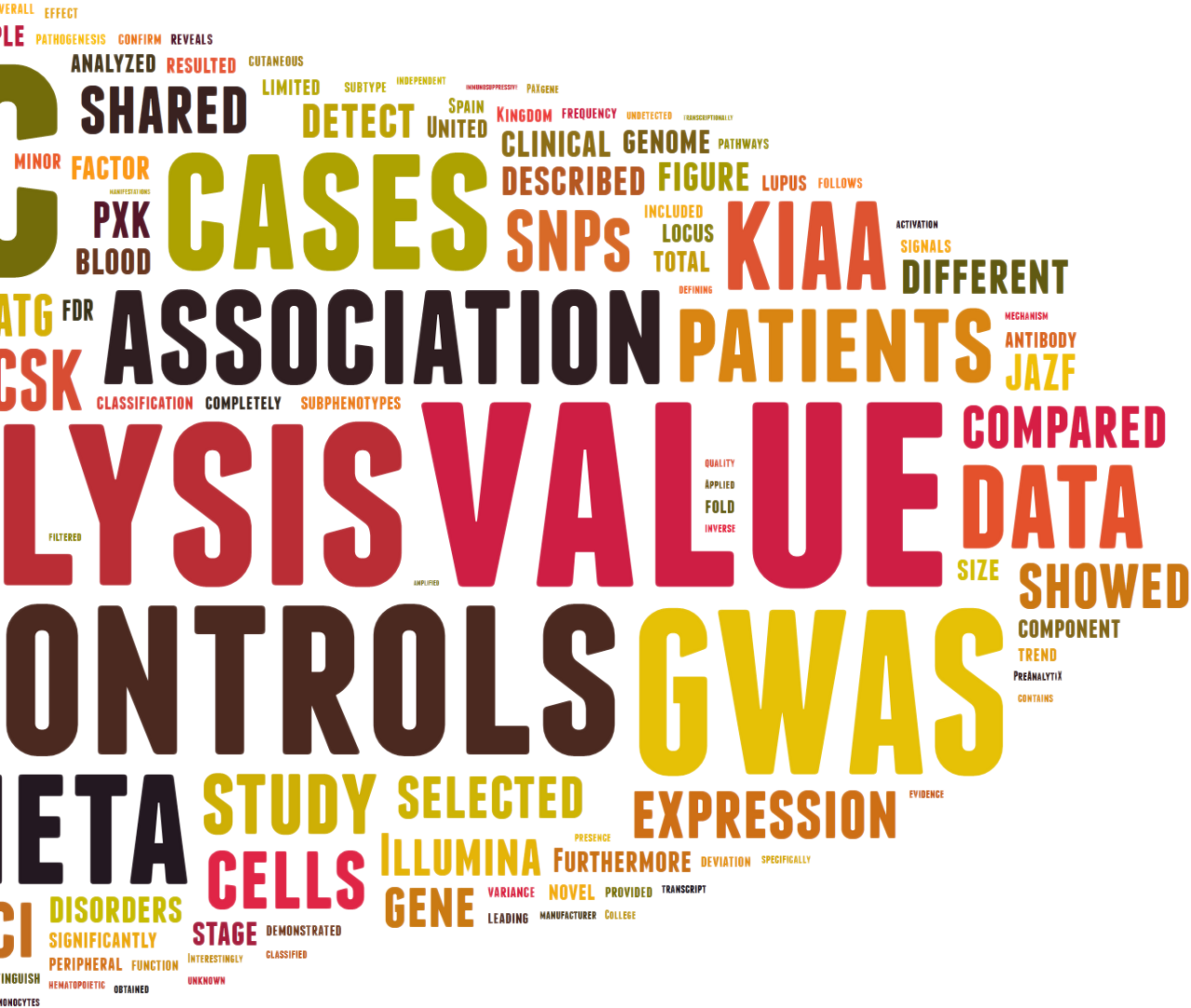
Supplementary Figure 1.



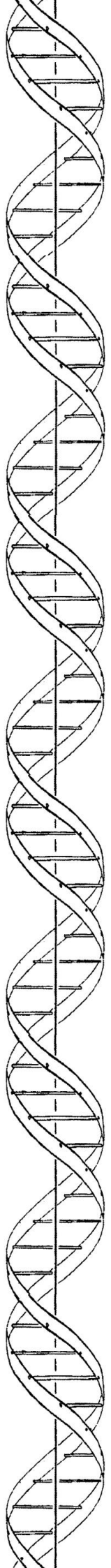
Supplementary Figure 2.







SYSTEMIC SCLEROSIS AND SYSTEMIC LUPUS
 ERYTHEMATOSUS PAN-META-GWAS REVEALS SIX
 NEW SHARED SUSCEPTIBILITY LOCI. UNDER REVIEW.

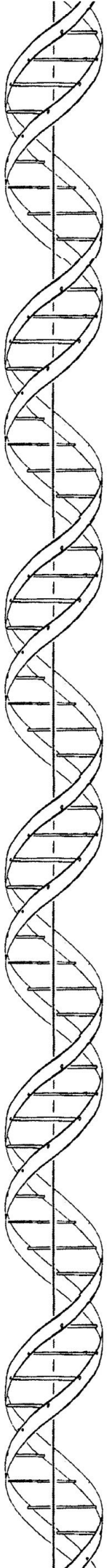


Systemic sclerosis and systemic lupus erythematosus pan-meta-GWAS reveals six new shared susceptibility loci

Jose-Ezequiel Martin^{1,36}, Shervin Assassi^{2,36}, Lina-Marcela Diaz-Gallo¹, Jasper C. Broen^{3, 4}, Carmen P. Simeon⁵, Luis Rodriguez-Rodriguez⁶, Esther Vicente-Rabaneda⁷, Vicente Fonollosa⁵, Norberto Ortego-Centeno⁸, Miguel A. González-Gay⁹, Paloma García de la Peña¹⁰, Patricia Carreira¹¹, Spanish Scleroderma Group¹², SLEGEN consortium¹³, U.S. Scleroderma GWAS group¹⁴, BIOLUPUS¹⁵, Mayte Camps¹⁶, Jose M. Sabio¹⁷, Sandra D'Alfonso¹⁸, Madelon C. Vonk³, Alexandre E. Voskuyl¹⁹, Annemie J. Schuerwegh²⁰, Alexander Kreuter²¹, Torsten Witte²², Gabriella Riemekasten²³, Nicolas Hunzelmann²⁴, Paolo Airo²⁵, Lorenzo Beretta²⁶, Raffaella Scorza²⁶, Claudio Lunardi²⁷, Jacob van Laar²⁸, Meng May Chee²⁹, Jane Worthington³⁰, Arianne Herrick³⁰, Christopher Denton³¹, Carmen Fonseca³¹, Filemon K. Tan², Frank Arnett², Xiaodong Zhou², John D. Reveille², Olga Gorlova³², Bobby P.C. Koeleman³³, Timothy R.D.J. Radstake^{4,36}, Timothy Vyse^{34,36}, Maureen D. Mayes^{2,36}, Marta E. Alarcón-Riquelme^{35,36}, Javier Martin^{1,36}.

Affiliations

¹Instituto de Parasitología y Biomedicina Lopez-Neyra, CSIC, Granada, Spain. ²The University of Texas Health Science Center–Houston, Houston, Texas, USA. ³Department of Rheumatology, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands. ⁴Department of Rheumatology, Clinical Immunology and Translational Immunology, University Utrecht Medical Center, Utrecht, The Netherlands. ⁵Servicio de Medicina Interna, Hospital Valle de Hebron, Barcelona, Spain. ⁶Servicio de Reumatología, Hospital Clínico San Carlos, Madrid, Spain. ⁷Servicio de Reumatología, Hospital La Princesa, Madrid, Spain. ⁸Servicio de Medicina Interna,

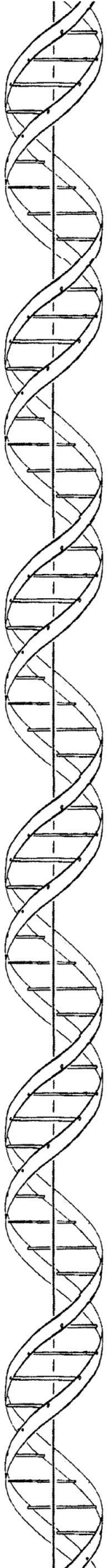


Hospital Clínico Universitario, Granada, Spain. ⁹Servicio de Reumatología, Hospital Marqués de Valdecilla, IFIMAV, Santander, Spain. ¹⁰Servicio de Reumatología, Hospital Ramón y Cajal, Madrid, Spain. ¹¹Hospital 12 de Octubre, Madrid, Spain. ¹²See *supplementary note 1*. ¹³see *supplementary note 2*. ¹⁴see *supplementary note 3*. ¹⁵See *supplementary note 4*. ¹⁶Servicio de Medicina Interna, Hospital Carlos Haya, Málaga, Spain. ¹⁷Unidad de Enfermedades Autoinmunes Sistémicas, Servicio de Medicina Interna, Hospital Universitario Virgen de las Nieves, Granada, Spain. ¹⁸Department of Medical Sciences and Institute of Research in Chronic Autoimmune Diseases (IRCAD), University of Eastern Piedmont, Novara, Italy. ¹⁹VU University Medical Center, Amsterdam, The Netherlands. ²⁰Department of Rheumatology, University of Leiden, Leiden, The Netherlands. ²¹Department of Dermatology, Josefs-Hospital, Ruhr University Bochum, Germany. ²²Hannover Medical School, Hannover, Germany. ²³Department of Rheumatology and Clinical Immunology, Charité University Hospital, Berlin, Germany. ²⁴Department of Dermatology, University of Cologne, Cologne, Germany. ²⁵Rheumatology Unit and Chair, Spedali Civili, Università de gli Studi, Brescia, Italy. ²⁶Referral Center for Systemic Autoimmune Diseases, Fondazione IRCCS Ca'Granda Ospedale Ma Repiore Policlinico and University of Milan, Milan, Italy. ²⁷Department of Medicine, Policlinico GB Rossi, University of Verona, Italy. ²⁸Institute of Cellular Medicine, Newcastle University, Newcastle Upon Tyne, United Kingdom. ²⁹Centre for Rheumatic Diseases, Glasgow Royal Infirmary Glasgow, UK. ³⁰Department of Rheumatology and Epidemiology, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK. ³¹Centre for Rheumatology, Royal Free and University College School, London. ³²Department of Epidemiology, M. D. Anderson Cancer Center, Houston, Texas, USA. ³³Department of Medical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands.

³⁴Divisions of Genetics and Molecular Medicine and Division of Immunology, Infection and Inflammatory Disease, King's College London, Guy's Hospital, London, UK.

³⁵Centro de Genómica e Investigación Oncológica (GENYO) Pfizer-Universidad de Granada-Junta de Andalucía, Granada, Spain. ³⁶These authors contributed equally to this work.

Keywords: Genetic susceptibility, meta-GWAS, autoimmunity, systemic sclerosis, systemic lupus erythematosus, *KIAA0319L*, *PXK*, *ATG5*, *JAZF1*, *SAMD9L*, *CSK*.



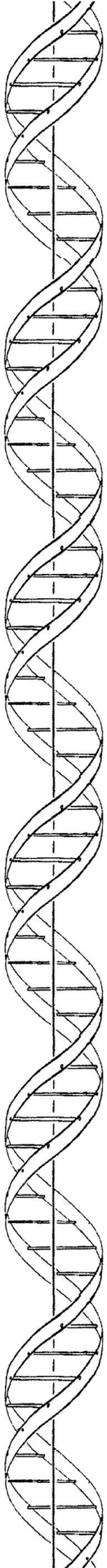
Abstract

Systemic sclerosis (SSc) and systemic lupus erythematosus (SLE) are two archetypal systemic autoimmune diseases which have been shown to share multiple genetic susceptibility loci. In order to gain insight into the genetic basis of these diseases we performed a pan-meta-analysis of two genome-wide association studies (GWAS) together with a replication stage including additional SSc and SLE cohorts. This increased the sample size to a total of 21,109 (6,835 cases and 14,274 controls). We selected for replication 20 SNPs from the GWAS data. We were able to validate as novel genetic susceptibility loci (for the combined SSc and SLE analysis) *KIAA0319L* ($P = 3.31 \times 10^{-11}$, OR = 1.49), *PXK* ($P = 3.27 \times 10^{-11}$, OR = 1.20), *ATG5* ($P = 5.30 \times 10^{-7}$, OR = 1.14), *JAZF1* ($P = 1.11 \times 10^{-8}$, OR = 1.13), *SAMD9L* ($P = 3.17 \times 10^{-7}$, OR = 1.19) and *CSK* ($P = 2.59 \times 10^{-7}$, OR = 1.13). Furthermore, we observed that *KIAA0319L* and *SAMD9L* were overexpressed in peripheral blood cells of SSc and SLE patients compared to healthy controls. With these, we add five (*KIAA0319L*, *PXK*, *ATG5*, *JAZF1* and *SAMD9L*) and three (*KIAA0319L*, *SAMD9L* and *CSK*) new susceptibility loci for SSc and SLE, respectively, increasing significantly our knowledge of the genetic basis of autoimmunity.

Introduction

Most autoimmune disorders are genetically complex and clinically heterogeneous. Classification permit physicians to distinguish individual autoimmune criteria based on typical clinical features. However, underlying these separate clinical features there is a genetic continuum of susceptibility factors and molecular pathways leading to autoimmunity. This becomes clear when analyzing existing genetic data according to clinical subphenotypes, reducing the overall genetic heterogeneity in the analyses (1-8). This has been clearly demonstrated in rheumatoid arthritis (RA) in which separating anti-citrullinated protein antibody (ACPA) positive and negative cases resulted in a far more homogeneous genetic analysis revealing genetic subgroups (1, 3-6). The aim of personalized medicine is thus to accomplish a degree of resolution that allows us to distinguish genetically and molecularly such groups and provide more targeted therapeutic interventions.

There is a growing body of evidence that all autoimmune disorders share to a varying degree their genetic susceptibility loci (9). It is clear that lack of statistical power due to limited sample size is a barrier to completely defining the genetic contribution to autoimmunity. As we increase our sample size, we can identify additional genes shared by some diseases but not shared by others. From the viewpoint that all autoimmune disorders are heterogeneous entities whose specific manifestations depend on the presence of several susceptibility genetic variants and environmental triggers, the combined analysis of different autoimmune disorders will greatly increase our statistical power to detect modest genetic effects shared among them. This approach has already been successfully used to detect the shared genetic susceptibility component of RA, celiac disease and Crohn's disease (10, 11). In the present study we meta-analyze the GWASs of two autoimmune disorders that have been demonstrated to share a relatively



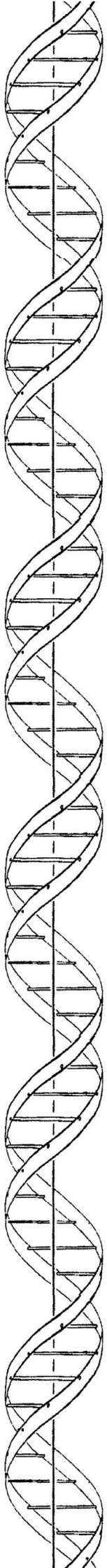
large portion of their genetic component and clinical features: systemic sclerosis and systemic lupus erythematosus (12, 13), identifying new loci for each disease. We would suggest a similar approach for other autoimmune diseases in the future.

Results

According to the selection criteria, we investigated 20 SNPs for replication whose statistics are shown in *table 1* and *figure 2*. According to the significance criteria established (see materials and methods), we could confirm one new association shared by SSc (more specifically in the limited cutaneous subtype) and SLE: rs2275247 in *KIAA0319L* (combined P value = 3.31×10^{-11} , OR = 1.49; SSc P value = 3.25×10^{-6} ; SLE P value = 1.15×10^{-5}) (*table 1*). Furthermore, we were able to determine two new genetic susceptibility regions at genome-wide level of significance in the combined analysis for SSc which had been previously described only in SLE: *PXK* (rs2176082, combined P value = 3.72×10^{-11} , OR = 1.20; SSc P value = 4.33×10^{-2} ; SLE P value = 2.06×10^{-5}) and *JAZF1* (combined P value = 1.11×10^{-8} , OR = 1.13; SSc P value = 1.27×10^{-2} ; SLE P value = 3.84×10^{-5}) but with a modest association particularly for the SSc analysis alone. This suggests a minor role of these genetic variants in SSc compared to SLE (*table 1*).

We also observed suggestive associations in three other loci: *ATG5*, *SAMD9L* and *CSK*. *ATG5* has been previously identified as a risk locus for SLE (14), but it is a novel susceptibility locus in SSc (rs3827644, combined P value = 5.30×10^{-7} , OR = 1.14; SSc P value = 8.09×10^{-3}). On the other hand, *CSK*, that has been previously described in SSc (15), showed a novel suggestive association for SLE (rs7172677, combined P value = 2.59×10^{-7} , OR = 1.13; SLE P value = 1.10×10^{-2}). *SAMD9L* is a completely novel locus for both diseases (rs1133906, combined P value = 3.17×10^{-7} , OR = 1.19; SSc P value = 1.65×10^{-1} ; SLE P value = 1.55×10^{-5}) (*table 1*).

In the whole blood gene expression data, we observed that *KIAA0319L* was significantly overexpressed in SLE patients compared to unaffected controls ($P = 5.36 \times 10^{-5}$, FDR = 2.94×10^{-3}) with a fold change of 1.9, while in SSc, there was a trend



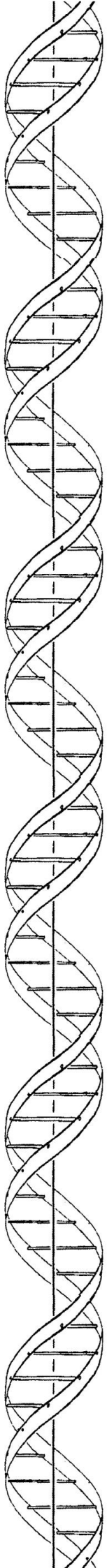
towards overexpression compared to healthy individuals (P value = 9.05×10^{-3} , FDR = 0.14) with a fold change of 1.3. *SAMD9L* was also significantly overexpressed in SLE patients compared to unaffected controls with a fold change of 2.4 ($P < 1 \times 10^{-7}$, FDR $< 1 \times 10^{-7}$). Similarly, *SAMD9L* showed a trend for higher expression in SSc patients compared to unaffected controls with a fold-change of 1.4 ($P = 1.5 \times 10^{-2}$, FDR = 0.17). No significant expression differences were observed for *PXK*, *ATG5*, *JAZF1* and *CSK* in either SSc or SLE compared to controls (*table 2*).

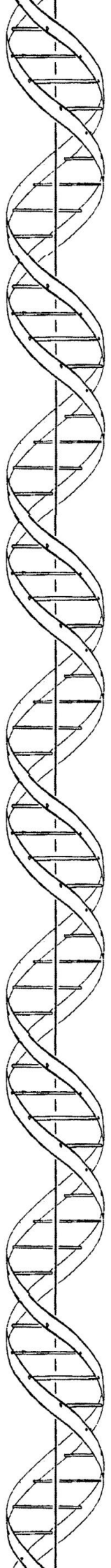
Discussion

Utilizing a GWAS pan-meta-analysis strategy, we were able to identify a total of six new autoimmunity susceptibility loci, five of which are new for SSc and three of which are new for SLE. These new genetic susceptibility loci were undetected in the previous SSc and SLE GWASs due to the lack of statistical power.

The clearest example of this is rs2275247 in *KIAA0319L*, with a minor allele frequency of the 3.43% in our GWAS cohorts, which would have gone undetected as *supplementary table 3* predicts. However, due to the combined analysis, we were able to capture this association in our meta-analysis. *KIAA0319L* was previously associated with learning and cognition disabilities (16). Interestingly, the protein contains among its evolutionary conserved domains a Polycystic Kidney Disease (PKD) domain, which is an immunoglobulin family-like domain of unclear function (17). Furthermore, *KIAA0319L* was significantly overexpressed in peripheral blood cells from SLE patients compared to those of healthy controls, also showed a similar trend was observed in SSc. Another possible lead pointing to the role of *KIAA0319L* in autoimmunity can be found in the expression profile of this gene in the bioGPS public database (<http://biogps.org/#goto=welcome>); although this gene is expressed ubiquitously it has a particularly high expression in immune cells such as macrophages, natural killer cells and other hematopoietic cells in the mouse (*supplementary figure 3*) and CD33+ myeloid cells and CD14+ monocytes in humans (*supplementary figure 4*).

SAMD9L (sterile alpha motif domain containing 9-like) is a gene of unknown function that has been described to be transcriptionally altered in response to type I interferons (18), which are known to play an important role in both diseases (19). Furthermore, as in the case of *KIAA0319L*, we found that *SAMD9L* was differentially expressed between





peripheral blood cells of SLE patients and healthy controls and showed a similar trend in SSc patients. The expression profile of this gene was consistent with its presence in different hematopoietic cell types such as 271 B lymphoblast cell line, CD8 and CD4 T cells, natural killer cells and monocytes in humans (*supplementary figure5*).

PXK, *ATG5* and *JAZF1* were previously described as SLE genetic risk factors (14, 20). Interestingly, in the present study we confirm all of them, but with a modest association with SSc. The role of these three genes in SLE pathogenesis remains largely unknown, it is also unclear whether their pathogenic mechanism is the same or different in SSc. Conversely, we found *CSK*, recently described to be associated with SSc (15), as a suggestive susceptibility factor in SLE. *CSK* plays a central role in the immune response as an inhibitor of *LYP* (encoded by *PTPN22*), which in turn, when dissociated from *CSK* inhibits T cell activation (21). The implication of *CSK* in SSc pathogenesis was suggested to be through its role in fibroblast activation and in skin fibrosis (22), although it needs to be further explored whether the role it has in SLE pathogenesis follows the same pathway or acts by another mechanism.

Due to the increased statistical power derived from the merging of two GWAS in SSc and SLE and the addition of large replication cohorts, we were able to establish two new functionally attractive susceptibility loci for SSc and SLE (*KIAA0319L* and *SAMD9L*), three new susceptibility loci for SSc (*ATG5*, *PXK* and *JAZF1*) and one new locus for SLE (*CSK*). This study, together with several others, adds evidence of the genetic continuum that underlies SSc, SLE and most autoimmune disorders.

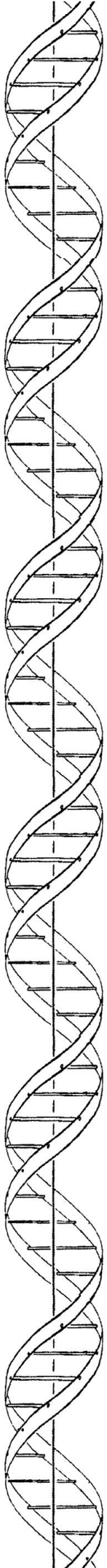
Material and Methods

Study cohorts

The GWAS cohorts analyzed in this study were composed of a total of 3,530 cases and 7,381 healthy controls. Of these 2,761 cases and 3,720 healthy controls belonged to a previous SSc GWAS conducted on cohorts from Spain, Germany, The Netherlands and USA (23). The rest of the cases and controls belonged to a previously published SLE GWAS conducted in the USA, SLEGEN cohort composed of 769 SLE patients and 3,661 healthy controls (14) (*supplementary tables 1 and 2*).

We selected independent SSc and SLE replication cohorts in order to confirm the results observed in the previous meta-GWAS stage. The SSc replication cohort was composed of 432 cases from Spain, 691 cases from Italy and 455 cases from the United Kingdom, while the SLE replication cohort was composed of 375 cases from Spain, 335 cases from Italy, and 1,017 cases from the United Kingdom. A shared set of independent controls was used to compare with the SSc and SLE patients composed of 760 controls from Spain, 481 controls from Italy and 5,652 controls from the United Kingdom (*supplementary table 1*). *Supplementary table 2* shows key features of the SSc cohorts analyzed. All the samples in the replication cohorts were recruited from hospitals and clinics of each country after approval by the corresponding ethics committees. Genotype data from the United Kingdom replication controls were obtained from the WTCCC repositories, for which we were granted access.

All cases either met the American College of Rheumatology preliminary criteria for the classification of SSc (24) or had at least three of the five CREST (calcinosis, Raynaud's phenomenon, esophageal dysmotility, sclerodactyly, telangiectasias) features and were classified according to their skin involvement and their auto-antibody production



status(25, 26). All SLE patients in the current study fulfilled the revised criteria for classification of SLE from the American College of Rheumatology (27). All individuals enrolled in the present study provided written informed consents.

Data quality control

GWAS data were filtered as previously described (23) using as a limits a 90% success call rate per SNP and individual, a deviation from Hardy-Weinberg equilibrium of a P value < 0.0001 and a minor allele frequency lower than 1%. The first ten principal components were estimated and individuals who deviated more than three standard deviations from the centroid of their population in the first two principal components were excluded as outliers. The replication cohorts were filtered in the same way except for the principal component analysis, which was not performed due to the lack of GWAS level data for these cohorts.

Genotyping

The GWAS genotyping of the SSc cases and controls was performed as follows: the Spanish SSc cases and controls together with Dutch and German SSc cases were genotyped at the Department of Medical Genetics of the University Medical Center Utrecht (The Netherlands) using the commercial release Illumina Human CNV370K BeadChip. Genotype data for Dutch and German controls were obtained from the Illumina Human 550K BeadChip available from a previous study. The SSc case group from the United States was genotyped at Boas Center for Genomics and Human Genetics, Feinstein Institute for Medical Research, North Shore Long Island Jewish Health System using the Illumina Human610-Quad BeadChip. CGEMS and Illumina iControlDB controls were genotyped on the Illumina Hap550K BeadChip. SNPs selected for the replication phase were genotyped in the replication cohorts using

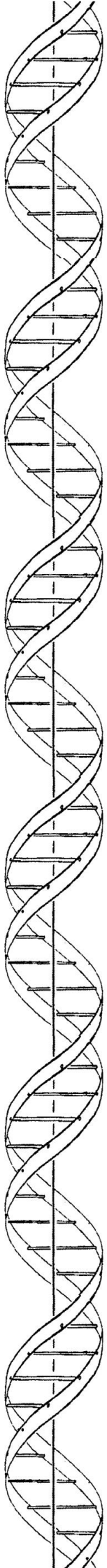
Applied Biosystems' TaqMan SNP assays on ABI Prism 7900 HT real-time thermocyclers.

Study design

Considering the size of our cohort for both study phases, simultaneously we calculated the statistical power for the different scenarios according to Skol *et al.* (28). *Supplementary table 3* shows the statistical power to detect different effect sizes. On average we had 86% statistical power to detect an OR of 1.20 with a minor allele frequency of 0.20.

Taking this into consideration, for the GWAS analysis three different criteria were followed to select SNPs for replication and maximization of our success in signal detection:

- 1) In order to detect common signals for SSc and SLE which caused either risk or protection for both diseases, we selected SNPs that showed a P value of the SSc and SLE meta-analysis $< 5 \times 10^{-7}$ and showed a nominally significant association with both diseases at P value < 0.05 , as well as no significant heterogeneity in the SSc cohorts meta-analysis ($Q > 0.05$) (seen in *supplementary table 4*).
- 2) To detect common signals for SSc and SLE which caused either risk or protection in one of the diseases and the opposite effect in the other, we selected SNPs that showed a P value of the SSc and SLE meta-analysis (using for SLE in this case $1/\text{OR}$ instead of the OR) $< 5 \times 10^{-7}$ and were associated with both a P value < 0.05 in both diseases, and without significant heterogeneity in the SSc cohorts meta-analysis ($Q > 0.05$) (seen in *supplementary table 5*).
- 3) To further detect any susceptibility variants previously reported for one of the diseases but not reported for the other, we selected any SNP with an overall P



value $< 5 \times 10^{-5}$ and associated separately in SSc and SLE (at a P value < 0.05) which had been previously described as a genetic risk factor for either disease (seen in *supplementary table 6*).

In any of these cases when performing the pan-meta-analysis with SSc, we also considered the most frequent SSc subphenotypes (classified as stated above): anti-centromere antibody (ACA) positive subgroup, anti-topoisomerase I antibody (ATA) positive subgroup, limited cutaneous subtype (lcSSc) and diffuse cutaneous subtype (dcSSc) (25, 26). *Supplementary figure 1* shows from which subphenotype of SSc each of the selected SNPs were derived from for the replication step.

After the replication stage was completed, we considered a signal to be statistically significant if the combined (SSc and SLE, GWAS and replication cohorts) meta-analysis P value was $< 5 \times 10^{-8}$, and if the meta-analysis P value (GWAS and replication cohorts) for each disease and stage was also significant. Furthermore, when a genetic variant presented a combined P value $< 5 \times 10^{-6}$ and was associated with a P value < 0.05 in the GWAS and replication stages' meta-analysis, it was considered a suggestive association.

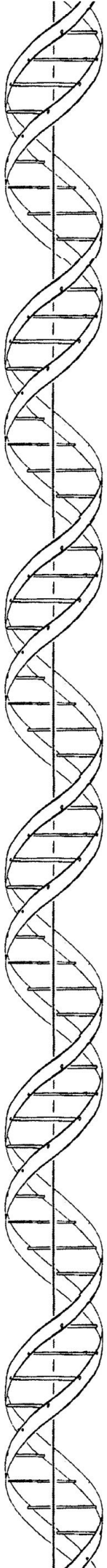
Statistical analysis

Meta-analysis of the SSc GWAS data was performed with the Cochran-Mantel-Haenszel test correcting for the genomic inflation factor lambda (GC correction) (*supplementary figure 2*). Analysis of the SLE GWAS data was performed by a simple χ^2 2x2 test correcting the P values for the genomic inflation factor lambda (GC correction) (*supplementary figure 2*). The pan-meta-analysis of the SSc and SLE GWAS data was performed with the inverse variance method calculating the resulting ORs. When the replication cohorts were included in the analysis, they were also

analyzed using the inverse variance method. When different associations were found in the same locus, the underlying association hit was determined by means of conditional logistic regression analysis. All statistical analyses were performed using Plink version 1.07 (29) (<http://pngu.mgh.harvard.edu/~purcell/plink/>) and HelixTree SNP Variation Suite 7 (<http://www.goldenhelix.com/>).

Gene Expression Data

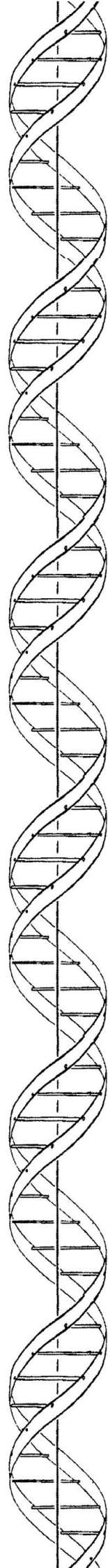
We had access to whole blood gene expression data from 74 SSc patients, 17 SLE patients and 21 healthy controls (19); which were not included in the GWAS cohorts. None of the patients were treated with immunosuppressive agents (exception prednisone ≤ 5 mg or hydroxychloroquine). Blood samples for transcript studies were drawn directly into PAXgene tubes (PreAnalytiX, Franklin Lakes, NJ). Total RNA was isolated according to the manufacturer's protocol using the PAXgene RNA kit (PreAnalytiX). The RNA quality and yield were assessed using a 2100 Bioanalyzer (Agilent, Palo Alto, CA) and an ND-1000 Spectrophotometer (NanoDrop Technologies, Wilmington, DE). Two hundred nanograms of total RNA was amplified and purified using the Illumina TotalPrep RNA Amplification Kit (Applied Biosystems/Ambion, Austin, TX) in accordance with the manufacturer's instructions. The amplified complementary RNA was hybridized on Illumina Human Ref-8 BeadChips, and the data were extracted with the Illumina Beadstudio software suite (Illumina, San Diego, CA). A transcript was defined as differentially expressed when the significance level for the comparison was $P \leq 0.05$ and the false discovery rate (FDR) was ≤ 0.10 using a random-variance t-test.



Acknowledgements and Funding

We thank Sofia Vargas, Sonia Garcia and Gema Robledo for their excellent technical assistance, and all the patients and healthy controls for kindly accepting their essential collaboration. We would also like to thank the following organizations: the EULAR Scleroderma Trials and Research (EUSTAR), the German Network of Systemic Sclerosis and Banco Nacional de ADN (University of Salamanca, Spain). This work was supported by the following grants: **J.M.** was funded by GEN-FER from the Spanish Society of Rheumatology, SAF2009-11110 and SAF2012-34435 from the Ministerio de Economía y Competitividad, CTS-4977 from Junta de Andalucía (Spain), Redes Temáticas de Investigación Cooperativa Sanitaria Program, RD08/0075 (RIER) from Instituto de Salud Carlos III (ISCIII, Spain), Fondo Europeo de Desarrollo Regional (FEDER) and the grant NIH NIAID 1U01AI09090-01. **T.R.D.J.R.** was funded by the VIDI laureate from the Dutch Association of Research (NWO) and Dutch Arthritis Foundation (National Reumafonds) and is an ERC starting grant Laureate 2011. **J.M.** and **T.R.D.J.R.** were sponsored by the Orphan Disease Program grant from the European League Against Rheumatism (EULAR). **B.P.C.K.** is supported by the Dutch Diabetes Research Foundation (grant 2008.40.001) and the Dutch Arthritis Foundation (Reumafonds, grant NR 09-1-408). **S.A.** and **M.D.M.** were supported by NIH/NIAMS Scleroderma Family Registry and DNA Repository (N01-AR-0-2251), NIH/NIAMS-RO1- AR055258, and NIH/NIAMS Center of Research Translation in Scleroderma (1P50AR054144), the Department of Defense Congressionally Directed Medical Research Programs (W81XWH-07-01-0111), and K23-AR-061436.. **M.E.A.R.** was funded from the Instituto de Salud Carlos III partially through FEDER funds of the European Union (PS09/00129), la Consejería de Salud de Andalucía (PI0012), la Fundación Ramón Areces and the Swedish Research Council. **M.E.A.R.** is Chairman of

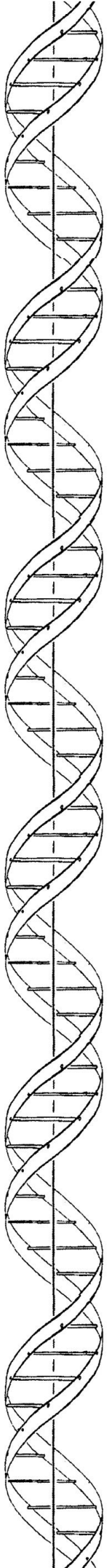
the BIOLUPUS network funded by the European Science Foundation. **T.W.** is funded by the grant KFO 250, TP03, WI 1031/6-1.



References

1. Terao, C., Ohmura, K., Ikari, K., Kochi, Y., Maruya, E., Katayama, M., Yurugi, K., Shimada, K., Murasawa, A., Honjo, S. *et al.* (2012) ACPA-Negative RA Consists of Two Genetically Distinct Subsets Based on RF Positivity in Japanese. *PLoS One*, **7**, e40067.
2. Gorlova, O., Martin, J.E., Rueda, B., Koeleman, B.P., Ying, J., Teruel, M., Diaz-Gallo, L.M., Broen, J.C., Vonk, M.C., Simeon, C.P. *et al.* (2011) Identification of novel genetic markers associated with clinical phenotypes of systemic sclerosis through a genome-wide association strategy. *PLoS Genet*, **7**, e1002178.
3. Lundberg, K., Bengtsson, C., Kharlamova, N., Reed, E., Jiang, X., Kallberg, H., Pollak-Dorocic, I., Israelsson, L., Kessel, C., Padyukov, L. *et al.* (2012) Genetic and environmental determinants for disease risk in subsets of rheumatoid arthritis defined by the anticitrullinated protein/peptide antibody fine specificity profile. *Ann Rheum Dis*.
4. Mackie, S.L., Taylor, J.C., Martin, S.G., Wordsworth, P., Steer, S., Wilson, A.G., Worthington, J., Emery, P., Barrett, J.H. and Morgan, A.W. (2012) A spectrum of susceptibility to rheumatoid arthritis within HLA-DRB1: stratification by autoantibody status in a large UK population. *Genes Immun*, **13**, 120-8.
5. de Vries, R.R., van der Woude, D., Houwing, J.J. and Toes, R.E. (2011) Genetics of ACPA-positive rheumatoid arthritis: the beginning of the end? *Ann Rheum Dis*, **70 Suppl 1**, i51-4.
6. Padyukov, L., Seielstad, M., Ong, R.T., Ding, B., Ronnelid, J., Seddighzadeh, M., Alfredsson, L. and Klareskog, L. (2011) A genome-wide association study suggests contrasting associations in ACPA-positive versus ACPA-negative rheumatoid arthritis. *Ann Rheum Dis*, **70**, 259-65.
7. Sanchez, E., Nadig, A., Richardson, B.C., Freedman, B.I., Kaufman, K.M., Kelly, J.A., Niewold, T.B., Kamen, D.L., Gilkeson, G.S., Ziegler, J.T. *et al.* (2011) Phenotypic associations of genetic susceptibility loci in systemic lupus erythematosus. *Ann Rheum Dis*, **70**, 1752-7.
8. Chung, S.A., Taylor, K.E., Graham, R.R., Nititham, J., Lee, A.T., Ortmann, W.A., Jacob, C.O., Alarcon-Riquelme, M.E., Tsao, B.P., Harley, J.B. *et al.* (2011) Differential genetic associations for systemic lupus erythematosus based on anti-dsDNA autoantibody production. *PLoS Genet*, **7**, e1001323.
9. Cho, J.H. and Gregersen, P.K. Genomics and the multifactorial nature of human autoimmune disease. *N Engl J Med*, **365**, 1612-23.
10. Festen, E.A., Goyette, P., Green, T., Boucher, G., Beauchamp, C., Trynka, G., Dubois, P.C., Lagace, C., Stokkers, P.C., Hommes, D.W. *et al.* (2011) A meta-analysis of genome-wide association scans identifies IL18RAP, PTPN2, TAGAP, and PUS10 as shared risk loci for Crohn's disease and celiac disease. *PLoS Genet*, **7**, e1001283.
11. Zhernakova, A., Stahl, E.A., Trynka, G., Raychaudhuri, S., Festen, E.A., Franke, L., Westra, H.J., Fehrmann, R.S., Kurzeeman, F.A., Thomson, B. *et al.* (2011) Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. *PLoS Genet*, **7**, e1002004.
12. Martin, J.E., Bossini-Castillo, L. and Martin, J. (2012) Unraveling the genetic component of systemic sclerosis. *Hum Genet*, **131**, 1023-37.
13. Delgado-Vega, A., Sanchez, E., Lofgren, S., Castillejo-Lopez, C. and Alarcon-Riquelme, M.E. Recent findings on genetics of systemic autoimmune diseases. *Curr Opin Immunol*, **22**, 698-705.
14. Harley, J.B., Alarcon-Riquelme, M.E., Criswell, L.A., Jacob, C.O., Kimberly, R.P., Moser, K.L., Tsao, B.P., Vyse, T.J., Langefeld, C.D., Nath, S.K. *et al.* (2008) Genome-wide

- association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXX, KIAA1542 and other loci. *Nat Genet*, **40**, 204-10.
15. Martin, J.E., Broen, J.C., Carmona, F.D., Teruel, M., Simeon, C.P., Vonk, M.C., van 't Slot, R., Rodriguez-Rodriguez, L., Vicente, E., Fonollosa, V. *et al.* (2012) Identification of CSK as a systemic sclerosis genetic risk factor through Genome Wide Association Study follow-up. *Hum Mol Genet*, **21**, 2825-35.
 16. Poelmans, G., Buitelaar, J.K., Pauls, D.L. and Franke, B. (2011) A theoretical molecular network for dyslexia: integrating available genetic findings. *Mol Psychiatry*, **16**, 365-82.
 17. Bycroft, M., Bateman, A., Clarke, J., Hamill, S.J., Sandford, R., Thomas, R.L. and Chothia, C. (1999) The structure of a PKD domain from polycystin-1: implications for polycystic kidney disease. *EMBO J*, **18**, 297-305.
 18. Pappas, D.J., Coppola, G., Gabatto, P.A., Gao, F., Geschwind, D.H., Oksenberg, J.R. and Baranzini, S.E. (2009) Longitudinal system-based analysis of transcriptional responses to type I interferons. *Physiol Genomics*, **38**, 362-71.
 19. Assassi, S., Mayes, M.D., Arnett, F.C., Gourh, P., Agarwal, S.K., McNearney, T.A., Chaussabel, D., Oommen, N., Fischbach, M., Shah, K.R. *et al.* (2010) Systemic sclerosis and lupus: points in an interferon-mediated continuum. *Arthritis Rheum*, **62**, 589-98.
 20. Gateva, V., Sandling, J.K., Hom, G., Taylor, K.E., Chung, S.A., Sun, X., Ortmann, W., Kosoy, R., Ferreira, R.C., Nordmark, G. *et al.* (2009) A large-scale replication study identifies TNIP1, PRDM1, JAZF1, UHRF1BP1 and IL10 as risk loci for systemic lupus erythematosus. *Nat Genet*, **41**, 1228-33.
 21. Vang, T., Liu, W.H., Delacroix, L., Wu, S., Vasile, S., Dahl, R., Yang, L., Musumeci, L., Francis, D., Landskron, J. *et al.* LYP inhibits T-cell activation when dissociated from CSK. *Nat Chem Biol*, **8**, 437-46.
 22. Skhirtladze, C., Distler, O., Dees, C., Akhmetshina, A., Busch, N., Venalis, P., Zwerina, J., Spriewald, B., Pileckyte, M., Schett, G. *et al.* (2008) Src kinases in systemic sclerosis: central roles in fibroblast activation and in skin fibrosis. *Arthritis Rheum*, **58**, 1475-84.
 23. Radstake, T.R., Gorlova, O., Rueda, B., Martin, J.E., Alizadeh, B.Z., Palomino-Morales, R., Coenen, M.J., Vonk, M.C., Voskuyl, A.E., Schuerwegh, A.J. *et al.* (2010) Genome-wide association study of systemic sclerosis identifies CD247 as a new susceptibility locus. *Nat Genet*, **42**, 426-9.
 24. (1980) Preliminary criteria for the classification of systemic sclerosis (scleroderma). Subcommittee for scleroderma criteria of the American Rheumatism Association Diagnostic and Therapeutic Criteria Committee. *Arthritis Rheum*, **23**, 581-90.
 25. LeRoy, E.C., Black, C., Fleischmajer, R., Jablonska, S., Krieg, T., Medsger, T.A., Jr., Rowell, N. and Wollheim, F. (1988) Scleroderma (systemic sclerosis): classification, subsets and pathogenesis. *J Rheumatol*, **15**, 202-5.
 26. LeRoy, E.C. and Medsger, T.A., Jr. (2001) Criteria for the classification of early systemic sclerosis. *J Rheumatol*, **28**, 1573-6.
 27. Hochberg, M.C. (1997) Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus. *Arthritis Rheum*, **40**, 1725.
 28. Skol, A.D., Scott, L.J., Abecasis, G.R. and Boehnke, M. (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet*, **38**, 209-13.
 29. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, **81**, 559-75.



Legends to figures

Figure 1. Workflow diagram showing the overall process followed during the present work.

Figure 2. Twin Manhattan plot representing the results of the SSc (left side, in blue), SLE (right side, in green) and combined GWAS analysis (in both sides, in grey). *SSc plotting represents either the total disease or any of its considered subphenotypes, *i.e.* ACA positive, ATA positive, lcSSc and dcSSc. ***ICAI* and *JAZF1* SNPs were selected according to selection criteria three (see materials and methods), and not because of reaching the significance threshold in the GWAS stage.

Table 1. Results of the GWAS analysis, replication analysis and combined analysis of the 20 SNPs selected for replication under different criteria (see materials and methods). *All SSc P values are referred to the specified SSc phenotype. Rows in light green represent replicated SNPs at suggestive level while dark green ones represent replicated SNPs at GWAS level.

Chr.	SNP	Change	Locus	SSc Group*	GWAS						Replication						Combined					
					SSc		SLE		Meta-analysis		SSc		SLE		Meta-analysis		SSc		SLE		Meta-analysis	
					P*	OR	P	OR	P	OR	P*	OR	P	OR	P	OR	P*	OR	P	OR	P	OR
1	rs2275247	C/T	<i>KIAA0319L</i>	lcSSc	8.73E-05	1.48	9.76E-04	1.71	4.48E-07	1.54	2.06E-05	1.62	1.91E-03	1.40	9.82E-06	1.45	3.25E-06	1.45	1.15E-05	1.49	3.31E-11	1.49
2	rs4907310	T/C	<i>ITPR1PL1</i>	dcSSc	3.71E-04	1.22	7.62E-05	0.79	1.17E-07	1.25	3.91E-02	1.10	9.99E-01	1.00	8.53E-01	1.01	8.04E-03	1.10	2.56E-02	0.93	2.29E-04	1.11
2	rs12466487	C/T	<i>R3HDM1</i>	dcSSc	7.20E-04	1.25	7.92E-06	0.71	4.69E-08	1.32	2.56E-04	1.23	4.78E-07	1.31	3.73E-04	0.84	1.99E-05	1.20	1.10E-01	1.07	2.55E-01	1.04
2	rs10498070	C/A	<i>DNPEP</i>	lcSSc	6.92E-03	1.14	5.74E-07	0.70	4.83E-07	1.23	6.56E-01	1.02	7.23E-02	0.92	1.55E-01	1.05	2.06E-01	1.05	2.50E-05	0.85	8.60E-06	1.13
3	rs2176082	A/G	<i>PXK</i>	ACA	6.30E-06	1.30	4.11E-04	1.24	1.16E-08	1.27	2.42E-02	1.12	6.71E-03	1.13	1.35E-04	1.15	4.33E-02	1.08	2.06E-05	1.16	3.27E-11	1.20
4	rs6814708	C/T	<i>SORCS2</i>	ACA	3.15E-03	0.85	5.53E-06	1.30	1.30E-07	0.81	6.56E-01	0.98	6.16E-01	0.98	8.61E-01	0.99	2.73E-01	0.96	2.59E-02	1.08	2.94E-04	0.91
5	rs285912	G/C	<i>KCNV2</i>	ACA	7.71E-04	0.82	5.10E-05	1.28	1.61E-07	0.80	2.25E-01	0.94	5.32E-01	0.97	9.84E-01	1.00	7.08E-03	0.90	6.27E-02	1.07	4.71E-04	0.91
6	rs2145748	A/G	<i>CDC5L</i>	lcSSc	6.89E-04	0.84	1.30E-04	1.28	4.54E-07	0.82	5.74E-01	1.03	6.49E-01	0.98	5.37E-01	1.02	3.48E-01	0.96	6.00E-02	1.07	2.76E-03	0.92
6	rs3827644	C/G	<i>ATG5</i>	SSc	5.36E-03	1.13	1.73E-06	1.38	7.06E-07	1.20	5.30E-01	1.04	2.76E-02	1.11	3.27E-02	1.09	8.09E-03	1.13	4.66E-06	1.20	5.30E-07	1.14
7	rs4725072	A/C	<i>JGAI</i>	SSc	2.37E-03	1.27	2.88E-03	1.41	2.92E-05	1.31	1.30E-01	1.15	1.33E-01	1.13	3.69E-02	1.14	1.77E-02	1.18	3.44E-03	1.21	1.19E-05	1.21
7	rs1635852	T/C	<i>JAZF1</i>	SSc	1.72E-03	1.12	2.73E-03	1.18	2.08E-05	1.14	1.45E-02	1.12	3.42E-03	0.89	1.33E-04	1.12	1.27E-02	1.09	3.84E-05	1.14	1.11E-08	1.13
7	rs1133906	T/C	<i>SAMD9L</i>	lcSSc	5.12E-04	1.25	1.54E-05	1.43	7.45E-08	1.31	4.41E-01	1.05	2.41E-02	1.14	4.18E-02	1.10	1.65E-01	1.07	1.55E-05	1.23	3.17E-07	1.19
9	rs1002841	C/T	<i>SEC61B</i>	ACA	8.57E-03	1.16	1.53E-06	0.74	2.34E-07	1.24	8.56E-01	1.01	1.34E-01	1.06	1.84E-01	0.95	3.55E-01	1.03	1.44E-01	0.95	1.08E-02	1.06
9	rs7038399	A/G	<i>AKAP2</i>	dcSSc	4.19E-04	0.68	6.51E-05	0.64	1.21E-07	0.66	1.28E-01	1.12	7.69E-01	1.02	7.30E-01	1.02	5.78E-01	1.03	6.40E-02	0.90	4.81E-04	0.85
10	rs10903528	A/G	<i>ADARB2</i>	ATA	1.94E-05	1.51	2.61E-03	0.75	2.70E-07	1.42	7.26E-01	0.97	3.10E-01	0.94	3.78E-01	1.05	1.96E-01	1.07	1.19E-02	0.87	6.29E-05	1.19
11	rs1355223	A/G	<i>EHF</i>	lcSSc	2.42E-04	1.17	3.88E-04	1.22	3.98E-07	1.19	4.87E-02	1.10	5.41E-01	1.02	1.16E-01	1.05	3.17E-02	1.08	1.14E-02	1.09	4.36E-06	1.11
12	rs10161149	T/C	<i>NAI3</i>	dcSSc	9.96E-03	1.32	4.17E-07	1.86	2.00E-07	1.52	5.95E-01	1.05	7.39E-01	1.03	3.45E-01	1.07	9.74E-02	1.12	2.05E-03	1.24	7.05E-05	1.24
15	rs2925256	T/C	<i>UBE3A</i>	ATA	3.86E-06	2.34	6.11E-03	1.65	2.09E-07	1.97	7.71E-01	0.96	6.21E-01	0.94	4.82E-01	0.92	2.68E-01	1.13	2.41E-01	1.13	1.48E-02	1.24
15	rs9920771	T/C	<i>NEO1</i>	ATA	5.45E-03	0.55	1.49E-07	2.12	6.06E-09	0.50	5.29E-01	0.93	4.13E-01	1.08	5.20E-01	0.95	3.86E-02	0.84	5.13E-04	1.32	1.04E-04	0.76
15	rs7172677	A/C	<i>CSK</i>	SSc	2.08E-06	1.21	4.70E-03	1.19	3.38E-08	1.21	1.39E-01	1.08	2.63E-01	1.05	7.09E-02	1.06	1.92E-04	1.15	1.10E-02	1.10	2.59E-07	1.13

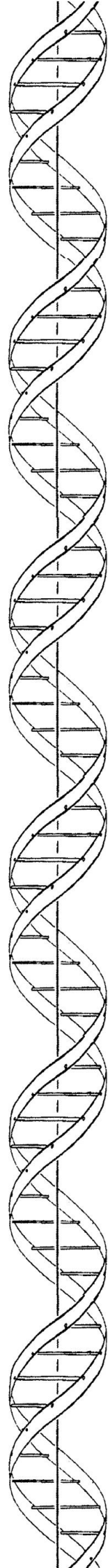


Table 2. Expression data of the successfully replicated loci from *table 1*, whether at GWAS level or suggestive level of association. Significantly overexpressed genes are marked in Bold. **PXK* and *ATG5* were not sufficiently expressed in whole blood samples and did not pass the filtering criteria.

Chr.	Gene	SSc Vs Controls			SLE Vs Controls		
		Parametric <i>P</i>	FDR	Fold change	Parametric <i>P</i>	FDR	Fold change
1	<i>KIAA0319L</i>	9.05E-03	1.43E-01	1.328	5.36E-05	2.94E-03	1.937
3	<i>PXK</i>	ND*	ND*	ND*	ND*	ND*	ND*
6	<i>ATG5</i>	ND*	ND*	ND*	ND*	ND*	ND*
7	<i>JAZF1</i>	9.04E-01	9.60E-01	0.990	3.90E-01	6.17E-01	0.870
7	<i>SAMD9L</i>	1.50E-02	1.70E-01	1.410	< 1E-07	< 1E-07	2.444
15	<i>CSK</i>	5.51E-01	7.70E-01	1.040	3.52E-01	5.80E-01	1.080

Figure 1.

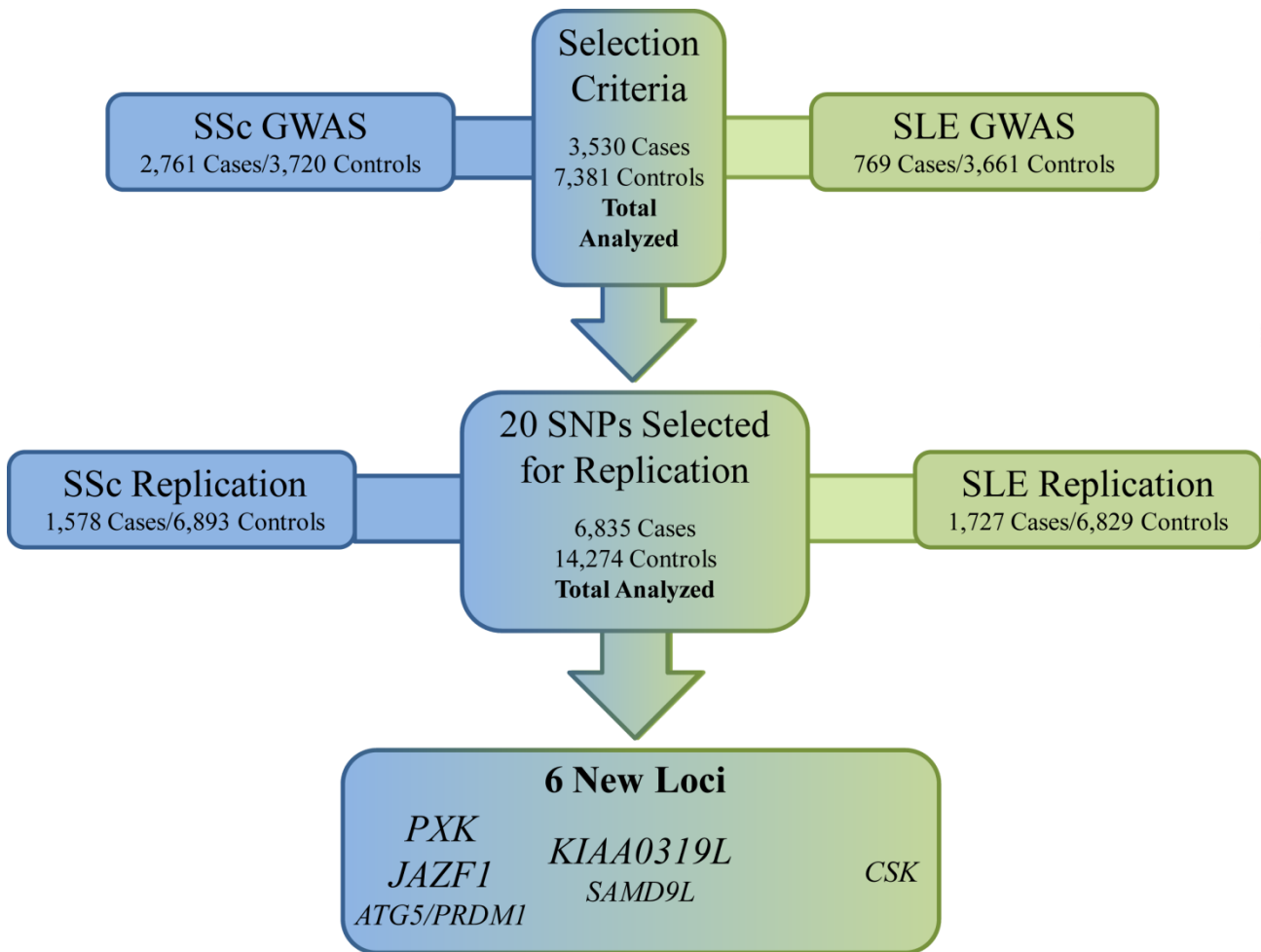
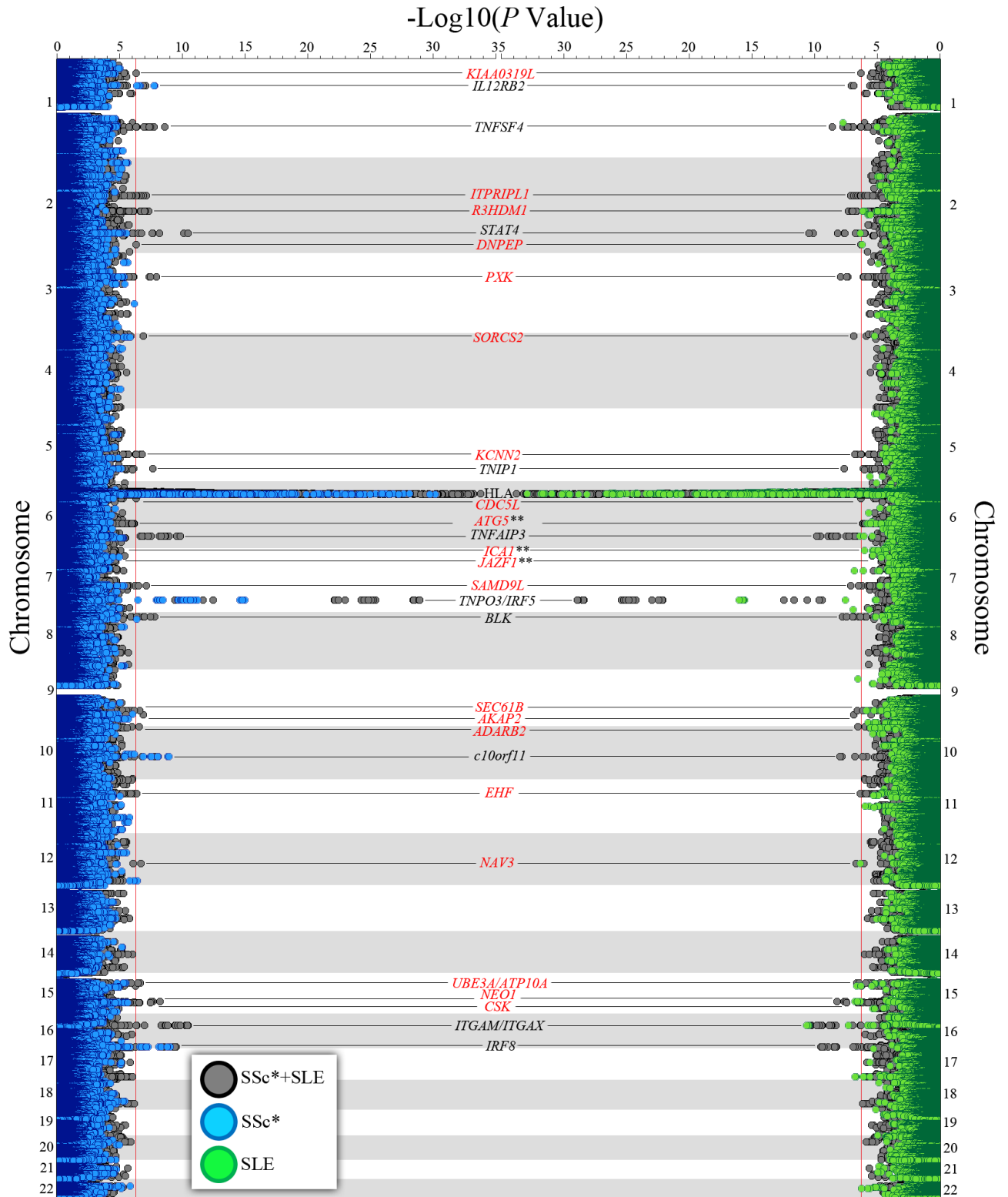
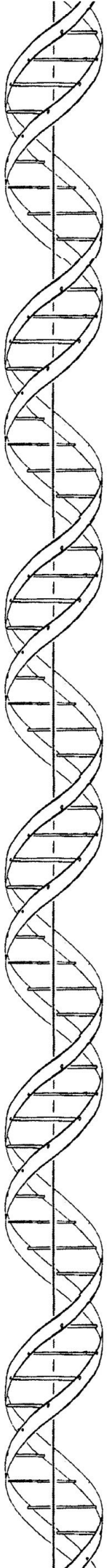


Figure 2.





Legends to supplementary figures

Supplementary figure 1. Manhattan plot broken down by SSc subphenotypes. In each case the analysis performed was as follows: SLE joined with the corresponding SSc phenotype using the inverse variance method.

Supplementary figure 2. QQ plots for SSc, SLE and the pan-meta-analysis of both, with or without the largely known and strongly associated MHC region. The lambdas observed were: SSc + SLE excluding the MHC region: 1.176; SSc excluding the MHC region: 1.246; SLE excluding the MHC region: 1.389; SSc + SLE including the MHC region: 1.195; SSc including the MHC region: 1.262; SLE including the MHC region: 1.414.

Supplementary figure 3. *KIAA0319L* expression profile in mice from the bioGPS database (<http://biogps.org/#goto=welcome>).

Supplementary figure 4. *KIAA0319L* expression profile in human from the bioGPS database (<http://biogps.org/#goto=welcome>).

Supplementary figure 5. *SAMD9L* expression profile in human from the bioGPS database (<http://biogps.org/#goto=welcome>).

Supplementary tables

Supplementary table 1. Study cohorts breakdown by the numbers. *These samples came from the GWAS in SSc by Radstake *et al.* †These samples came from the GWAS in SLE by Harley *et al.* ‡These samples came from the WTCCC shared UK control cohort. •These samples have been previously used and described in Radstake *et al.* and Gorlova *et al.* ΩThese samples came from Bentham & Morris, *et al.*, manuscript submitted for publication.

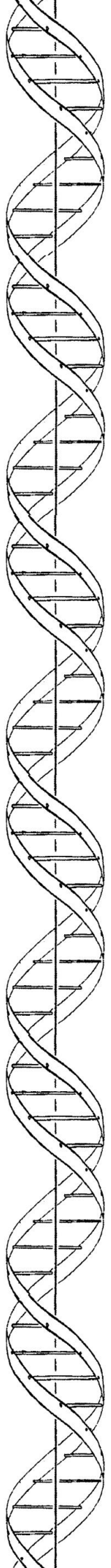
		SSc	SLE	Cases	Controls	Total
	Combined	4,339	2,496	6,835	14,274	21,109
	All	2,761	769	3,530	7,381	10,911
GWAS	Spain	781*	0	781	936*	1,717
	Germany	285*	0	285	670*	955
	Netherlands	203*	0	203	643*	846
	US	1,492*	769†	2,261	5,132*,†	7,393
	All	1,578	1,727	3,305	6,893	10,198
Replication	Spain	432•	375Ω	807	760•	1,567
	Italy	691•	335Ω	1,026	481•	1,507
	UK	455•	1,017Ω	1,472	5652‡	7,124

Supplementary table 2. Key features of the SSc GWAS and replication cohorts.

Population	Population Size		Sex (cases/controls)		Subtype		ACA Positive	ATA Positive
	Cases	Controls	Female	Male	Diffuse	Limited		
Overall	4,339	10,613	0.86/0.75	0.14/0.25	0.33	0.67	0.37	0.24
GWAS SSc cohorts								
Spain	781	936	0.90/0.73	0.10/0.27	0.30	0.70	0.50	0.26
Germany	285	670	0.88/0.62	0.12/0.38	0.43	0.57	0.45	0.33
The Netherlands	203	643	0.72/0.51	0.28/0.49	0.24	0.76	0.25	0.28
US	1,492	1,471	0.88/0.88	0.12/0.12	0.36	0.64	0.32	0.17
Replication SSc cohorts								
Spain	432	760	0.84/0.66	0.16/0.34	0.30	0.70	0.37	0.23
Italy	691	481	0.91/0.51	0.09/0.49	0.25	0.75	0.43	0.38
UK	455	5,652	0.83/0.81	0.17/0.19	0.29	0.71	0.36	0.18

Supplementary table 3. Power calculations for the present study considering the most probable scenarios in both the combined analysis and the separate diseases.

Group	N		Level of Significance	OR 1.30						OR 1.20					
	Cases	Controls		MAF	MAF	MAF	MAF	MAF	MAF	MAF	MAF	MAF	MAF	MAF	MAF
				0.40	0.30	0.20	0.15	0.10	0.05	0.40	0.30	0.20	0.15	0.10	0.05
SSc+SLE	6,835	14,274	5×10^{-8}	100	100	100	100	100	92	99	98	86	64	27	2
SSc	4,339	10,613	5×10^{-8}	100	100	100	100	100	63	91	82	54	30	8	0
SLE	2,496	10,554	5×10^{-8}	100	100	100	99	83	16	50	36	15	6	1	0



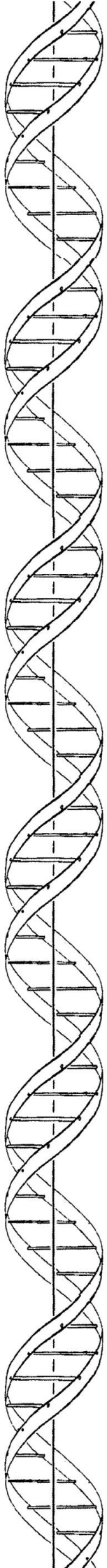
Supplementary table 4. All SNPs with an SSc and SLE combined meta-analysis P value lower than 5×10^{-6} in either of the following analyses: SSc + SLE, lcSSc + SLE, dcSSc + SLE, ACA positive SSc + SLE or ATA positive SSc + SLE. For each associated region only the independent signals, according to conditional logistic regression analyses are shown. For each SNP only the analysis group in which it was most significant is shown. SNPs selected for replication in this analysis are marked in **bold**. *Meta-analysis P value for SSc or its considered subgroup. ‡*TNFAIP3* presented three different signals in the SSc, dcSSc and ATA positive subgroups. **Due to its size, this table can be found in the attached excel file as supplementary information.**

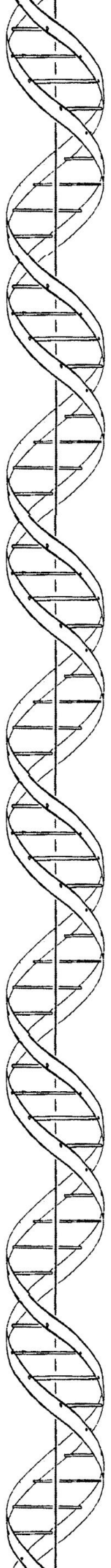
Supplementary table 5. All SNPs with an SSc and SLE inverted OR combined meta-analysis P value lower than 5×10^{-6} in either of the following analyses: SSc + SLE, lcSSc + SLE, dcSSc + SLE, ACA positive SSc + SLE or ATA positive SSc + SLE. For each associated region only the independent signals, according to conditional logistic regression analyses are shown. For each SNP only the analysis group in which it was most significant is shown. SNPs selected for replication in this analysis are marked in **bold**. *Meta-analysis P value for SSc or its considered subgroup. **Due to its size, this table can be found in the attached excel file as supplementary information.**

Supplementary table 6. Previously described associations in SSc, SLE or both of them analyzed in our data. SNPs only previously associated with one of the disorders which presented a combined P value lower than 5×10^{-5} and a SSc and SLE P value lower than 0.05 were selected for further replication. SNPs selected for replication in this analysis are marked in **bold**. *Supplementary references are found at the end of the Supplementary Material. **Due to its size, this table can be found in the attached excel file as supplementary information.**

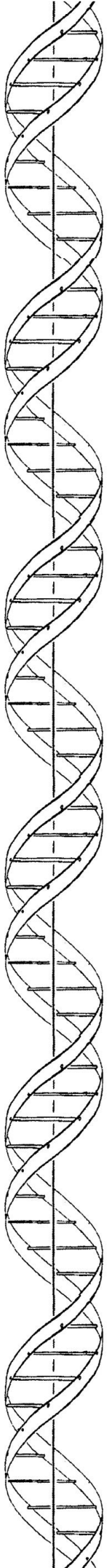
Supplementary References (as numbered in *supplementary table 6*)

1. Bossini-Castillo L, Martin JE, Broen J, Gorlova O, Simeon CP, Beretta L, et al. A GWAS follow-up study reveals the association of the IL12RB2 gene with systemic sclerosis in Caucasian populations. *Hum Mol Genet.* 2012;21(4):926-33.
2. Gourh P, Tan FK, Assassi S, Ahn CW, McNearney TA, Fischbach M, et al. Association of the PTPN22 R620W polymorphism with anti-topoisomerase I- and anticentromere antibody-positive systemic sclerosis. *Arthritis Rheum.* 2006;54(12):3945-53.
3. Kyogoku C, Langefeld CD, Ortmann WA, Lee A, Selby S, Carlton VE, et al. Genetic association of the R620W polymorphism of protein tyrosine phosphatase PTPN22 with human SLE. *Am J Hum Genet.* 2004;75(3):504-7.
4. Pricop L, Li L, Salmon JE, Jacob CO. Characterization of the FcγRIIIA promoter and 5'UTR sequences in patients with systemic lupus erythematosus. *Genes Immun.* 2002;3 Suppl 1:S47-50.
5. Radstake TR, Gorlova O, Rueda B, Martin JE, Alizadeh BZ, Palomino-Morales R, et al. Genome-wide association study of systemic sclerosis identifies CD247 as a new susceptibility locus. *Nat Genet.* 2010;42(5):426-9.
6. Gorman CL, Russell AI, Zhang Z, Cunninghame Graham D, Cope AP, Vyse TJ. Polymorphisms in the CD3Z gene influence TCRζ expression in systemic lupus erythematosus patients and healthy controls. *J Immunol.* 2008;180(2):1060-70.
7. Gourh P, Arnett FC, Tan FK, Assassi S, Divecha D, Paz G, et al. Association of TNFSF4 (OX40L) polymorphisms with susceptibility to systemic sclerosis. *Ann Rheum Dis.* 69(3):550-5.
8. Cunninghame Graham DS, Graham RR, Manku H, Wong AK, Whittaker JC, Gaffney PM, et al. Polymorphism at the TNF superfamily gene TNFSF4 confers susceptibility to systemic lupus erythematosus. *Nat Genet.* 2008;40(1):83-9.
9. Harley JB, Alarcon-Riquelme ME, Criswell LA, Jacob CO, Kimberly RP, Moser KL, et al. Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXX1, KIAA1542 and other loci. *Nat Genet.* 2008;40(2):204-10.
10. Crilly A, Hamilton J, Clark CJ, Jardine A, Madhok R. Analysis of the 5' flanking region of the interleukin 10 gene in patients with systemic sclerosis. *Rheumatology (Oxford).* 2003;42(11):1295-8.
11. Gateva V, Sandling JK, Hom G, Taylor KE, Chung SA, Sun X, et al. A large-scale replication study identifies TNIP1, PRDM1, JAZF1, UHRF1BP1 and IL10 as risk loci for systemic lupus erythematosus. *Nat Genet.* 2009;41(11):1228-33.
12. Allanore Y, Saad M, Dieude P, Avouac J, Distler JH, Amouyel P, et al. Genome-wide scan identifies TNIP1, PSORS1C1, and RHOB as novel risk loci for systemic sclerosis. *PLoS Genet.* 7(7):e1002091.
13. Rueda B, Broen J, Simeon C, Hesselstrand R, Diaz B, Suarez H, et al. The STAT4 gene influences the genetic predisposition to systemic sclerosis phenotype. *Hum Mol Genet.* 2009;18(11):2071-7.
14. Remmers EF, Plenge RM, Lee AT, Graham RR, Hom G, Behrens TW, et al. STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus. *N Engl J Med.* 2007;357(10):977-86.

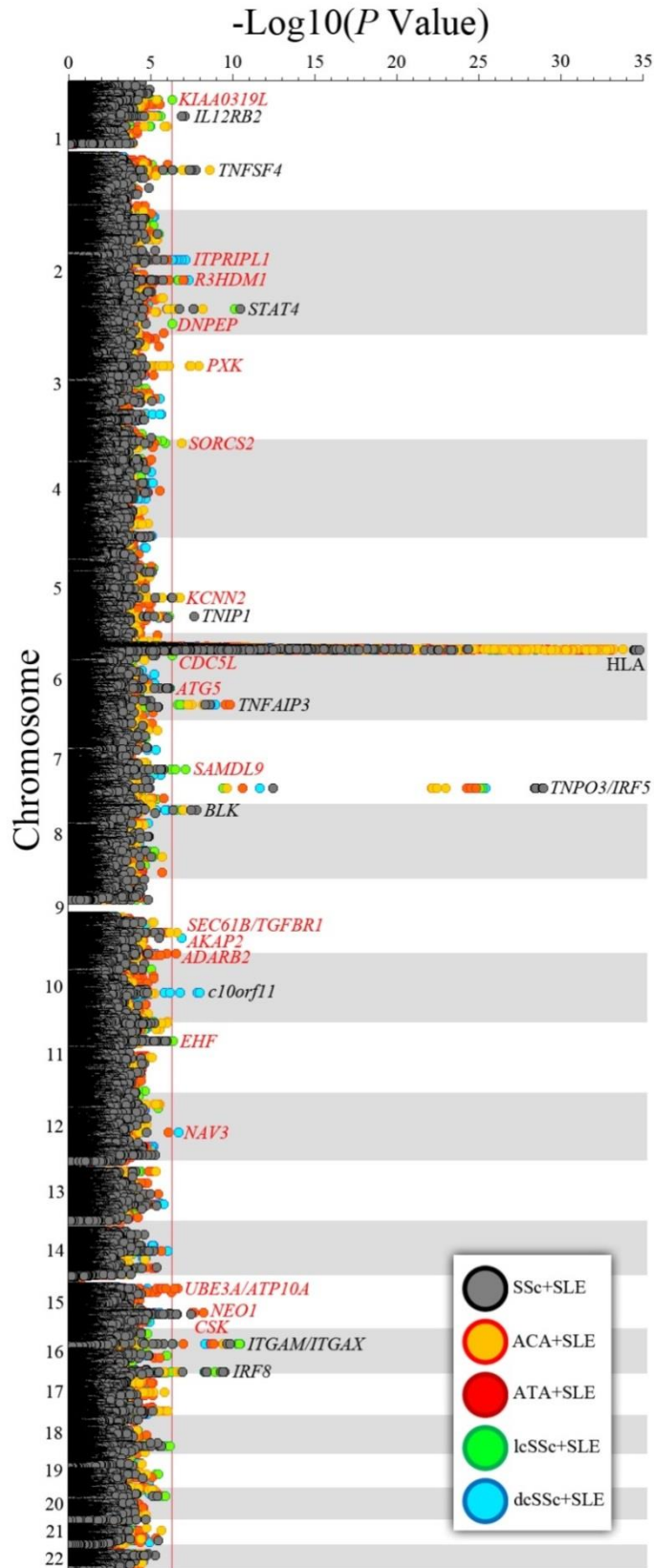


- 
15. Lessard CJ, Adrianto I, Ice JA, Wiley GB, Kelly JA, Glenn SB, et al. Identification of IRF8, TMEM39A, and IKZF3-ZPBP2 as susceptibility loci for systemic lupus erythematosus in a large-scale multiracial replication study. *Am J Hum Genet.*90(4):648-60.
 16. Dieude P, Wipff J, Guedj M, Ruiz B, Melchers I, Hachulla E, et al. BANK1 is a genetic risk factor for diffuse cutaneous systemic sclerosis and has additive effects with IRF5 and STAT4. *Arthritis Rheum.* 2009;60(11):3447-54.
 17. Kozyrev SV, Abelson AK, Wojcik J, Zaghlool A, Linga Reddy MV, Sanchez E, et al. Functional variants in the B-cell gene BANK1 are associated with systemic lupus erythematosus. *Nat Genet.* 2008;40(2):211-6.
 18. Dieude P, Guedj M, Wipff J, Ruiz B, Riemekasten G, Matucci-Cerinic M, et al. Association of the TNFAIP3 rs5029939 variant with systemic sclerosis in the European Caucasian population. *Ann Rheum Dis.*69(11):1958-64.
 19. Kawasaki A, Ito I, Ito S, Hayashi T, Goto D, Matsumoto I, et al. Association of TNFAIP3 polymorphism with susceptibility to systemic lupus erythematosus in a Japanese population. *J Biomed Biotechnol.*2010:207578.
 20. Dieude P, Guedj M, Wipff J, Avouac J, Fajardy I, Diot E, et al. Association between the IRF5 rs2004640 functional polymorphism and systemic sclerosis: a new perspective for pulmonary fibrosis. *Arthritis Rheum.* 2009;60(1):225-33.
 21. Sigurdsson S, Nordmark G, Goring HH, Lindroos K, Wiman AC, Sturfelt G, et al. Polymorphisms in the tyrosine kinase 2 and interferon regulatory factor 5 genes are associated with systemic lupus erythematosus. *Am J Hum Genet.* 2005;76(3):528-37.
 22. Gourh P, Agarwal SK, Martin E, Divecha D, Rueda B, Bunting H, et al. Association of the C8orf13-BLK region with systemic sclerosis in North-American and European populations. *J Autoimmun.* 2010;34(2):155-62.
 23. Hom G, Graham RR, Modrek B, Taylor KE, Ortmann W, Garnier S, et al. Association of systemic lupus erythematosus with C8orf13-BLK and ITGAM-ITGAX. *N Engl J Med.* 2008;358(9):900-9.
 24. Liakouli V, Manetti M, Pacini A, Tolusso B, Fatini C, Toscano A, et al. The -670G>A polymorphism in the FAS gene promoter region influences the susceptibility to systemic sclerosis. *Ann Rheum Dis.* 2009;68(4):584-90.
 25. Horiuchi T, Nishizaka H, Yasunaga S, Higuchi M, Tsukamoto H, Hayashi K, et al. Association of Fas/APO-1 gene polymorphism with systemic lupus erythematosus in Japanese. *Rheumatology (Oxford).* 1999;38(6):516-20.
 26. Martin JE, Carmona FD, Broen JC, Simeon CP, Vonk MC, Carreira P, et al. The autoimmune disease-associated IL2RA locus is involved in the clinical manifestations of systemic sclerosis. *Genes Immun.*13(2):191-6.
 27. Carmona FD, Gutala R, Simeon CP, Carreira P, Ortego-Centeno N, Vicente-Rabaneda E, et al. Novel identification of the IRF7 region as an anticentromere autoantibody propensity locus in systemic sclerosis. *Ann Rheum Dis.*71(1):114-9.
 28. Fu Q, Zhao J, Qian X, Wong JL, Kaufman KM, Yu CY, et al. Association of a functional IRF7 variant with systemic lupus erythematosus. *Arthritis Rheum.*63(3):749-54.
 29. Gorlova O, Martin JE, Rueda B, Koeleman BP, Ying J, Teruel M, et al. Identification of novel genetic markers associated with clinical phenotypes of systemic sclerosis through a genome-wide association strategy. *PLoS Genet.* 2011;7(7):e1002178.

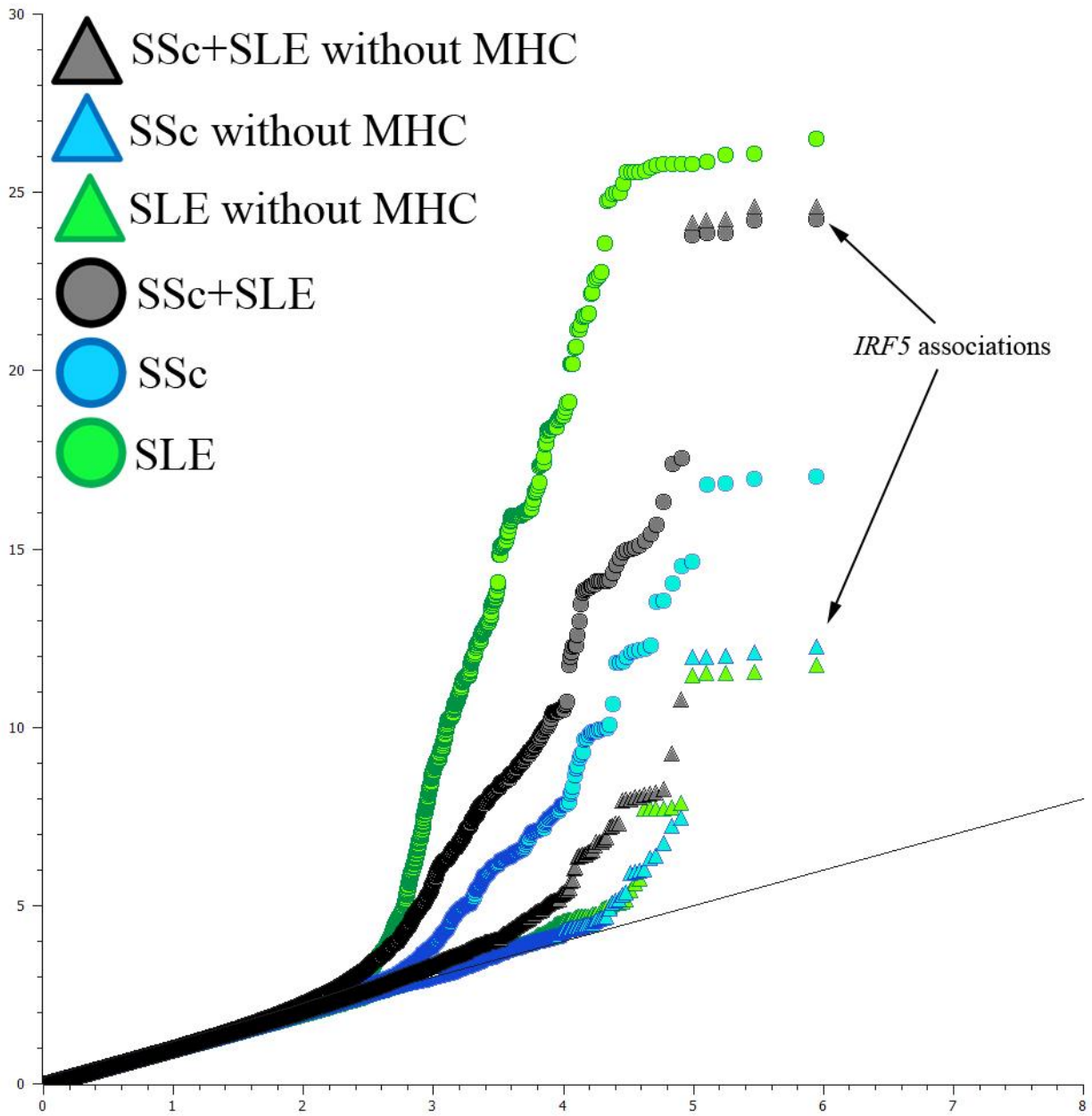
30. Martin JE, Broen JC, Carmona FD, Teruel M, Simeon CP, Vonk MC, et al. Identification of CSK as a systemic sclerosis genetic risk factor through Genome Wide Association Study follow-up. *Hum Mol Genet.* 2012;21(12):2825-35.
31. Carmona FD, Simeon CP, Beretta L, Carreira P, Vonk MC, Rios-Fernandez R, et al. Association of a non-synonymous functional variant of the ITGAM gene with systemic sclerosis. *Ann Rheum Dis.*70(11):2050-2.
32. Gourh P, Agarwal SK, Divecha D, Assassi S, Paz G, Arora-Singh RK, et al. Polymorphisms in TBX21 and STAT4 increase the risk of systemic sclerosis: evidence of possible gene-gene interaction and alterations in Th1/Th2 cytokines. *Arthritis Rheum.* 2009;60(12):3794-806.
33. Hasebe N, Kawasaki A, Ito I, Kawamoto M, Hasegawa M, Fujimoto M, et al. Association of UBE2L3 polymorphisms with diffuse cutaneous systemic sclerosis in a Japanese population. *Ann Rheum Dis.*71(7):1259-60.



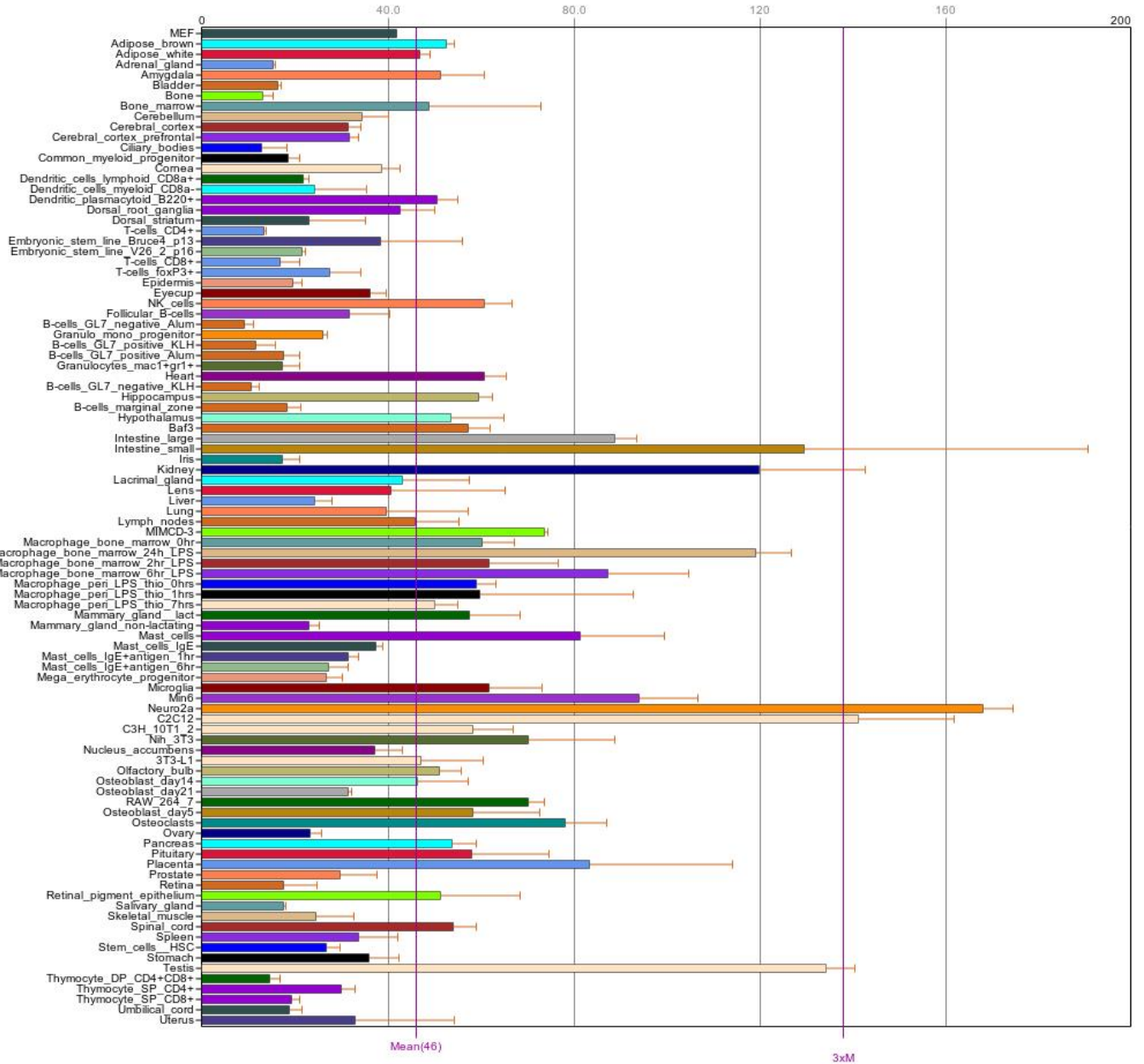
Supplementary Figure 1.



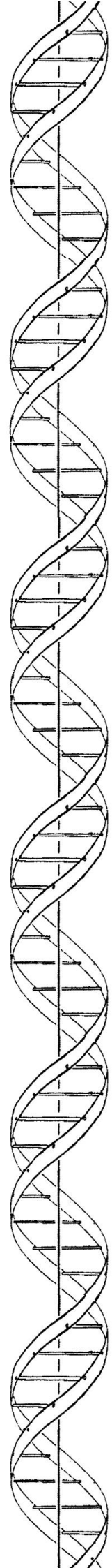
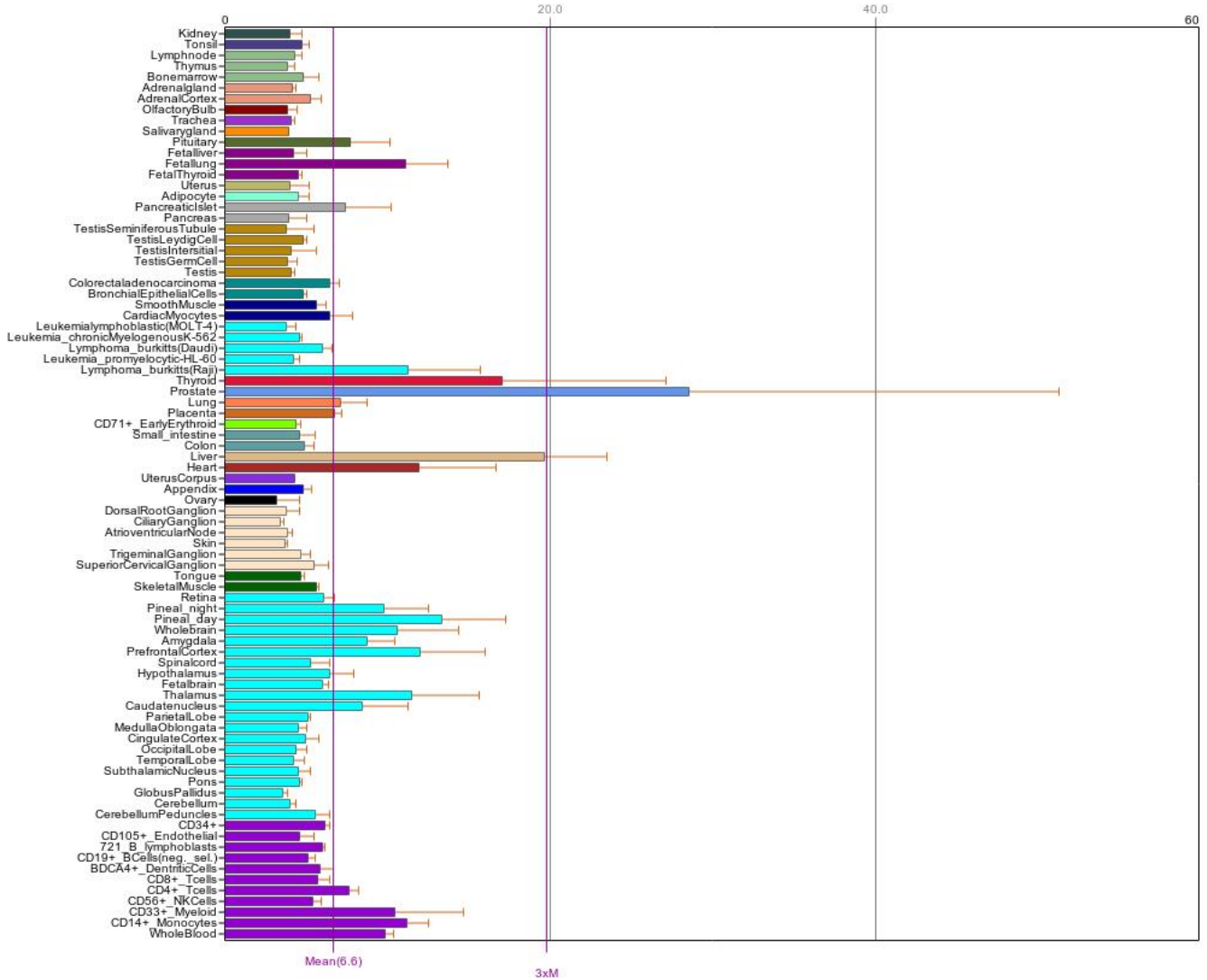
Supplementary Figure 2.



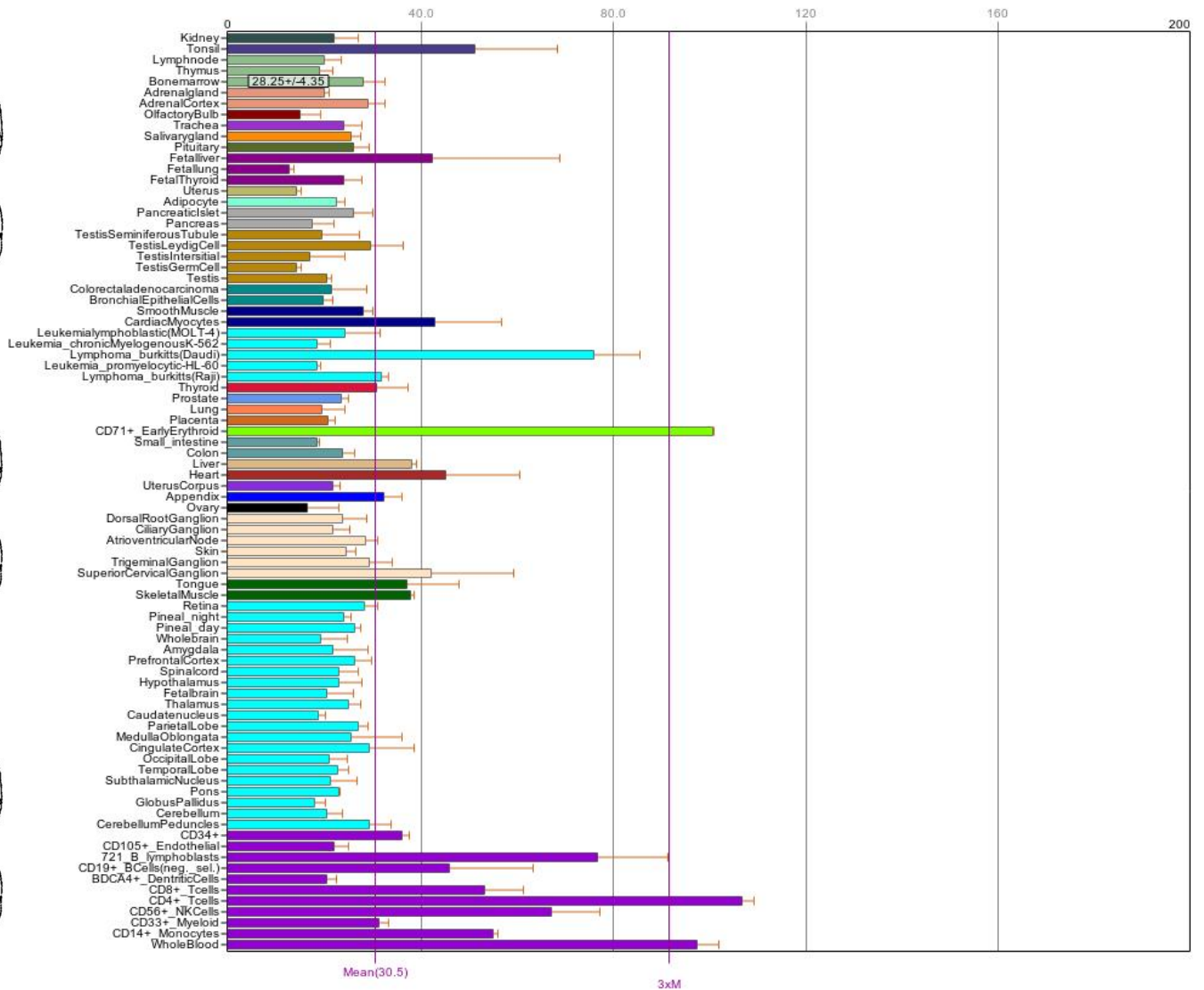
Supplementary Figure 3.

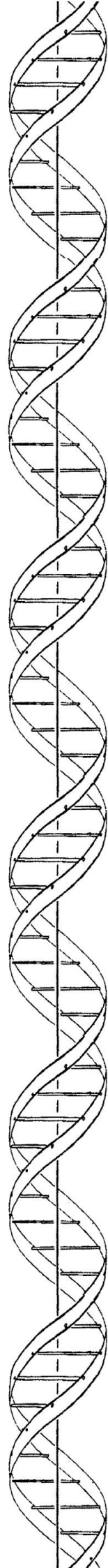


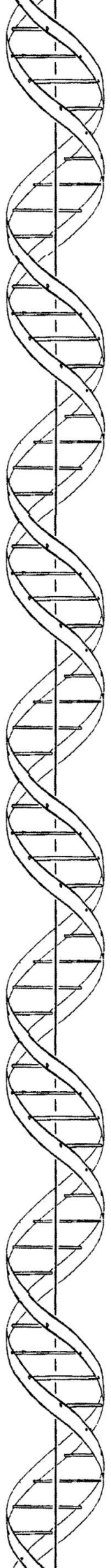
Supplementary Figure 4.



Supplementary Figure 5.



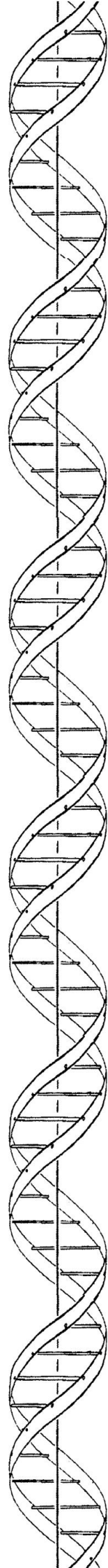




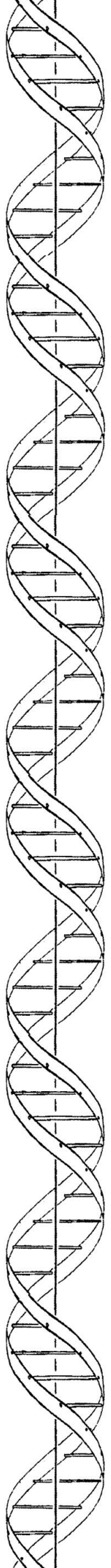
NOTHING SHOCKS ME. I'M A SCIENTIST.

-INDIANA JONES

DISCUSSION



DISCUSSION

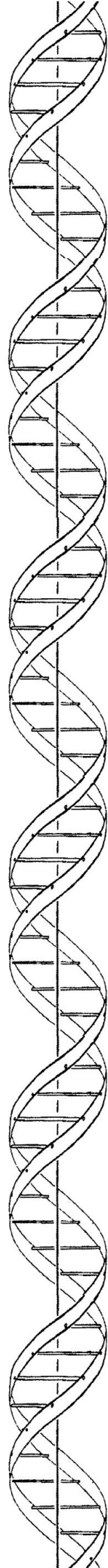


NEW GENETIC FINDINGS IN SYSTEMIC SCLEROSIS

Two GWAS have been performed in SSc in Caucasian populations to date, being the first one that of Radstake *et al.* presented in this thesis [45, 69]. In the GWAS performed by Allanore *et al.* three new genes were described as susceptibility genetic factors for SSc: *TNIP1*, *RHOB* and *PSORS1C1*. Of these three genes only *TNIP1* has been confirmed as a susceptibility factor for SSc, while *RHOB* has not been replicated and *PSORS1C1* was found to be dependent on the HLA class II genes association [69, 70]. In the GWAS presented in this thesis *CD247* was identified as a new susceptibility gene for SSc, association that has been confirmed by an independent study [45, 91]. In our SSc GWAS we also confirmed the previously described SSc loci *STAT4* and *IRF5* at GWAS level [45, 59, 62, 63].

In the SSc GWAS phenotype analysis performed during this thesis, we were able to determine *SOX5*, *GRB10*, *NOTCH4* and *IRF8* as novel susceptibility markers which confer risk towards the disease main phenotypes [92]. Furthermore in an exhaustive grey zone analysis of the GWAS data we were also able to determine that *CSK*, *NFKB1* and *PSD3* play an important role in the genetics of SSc [66]. Additionally, in this study we confirmed as SSc genetic risk loci the previously described genes *TNFSF4* and *TNFAIP3* [65-68]. Finally, using novel analysis techniques as pan-meta-GWAS and HLA molecules imputation we have been able to 1) determine that *JAZF1*, *KIAA0319L*, *PXK*, *ATG5* and *SAMD9L* are novel SSc risk factors shared with SLE, and 2) the association observed in the HLA region with SSc can be mostly explained with a seven aminoacid model in the HLA-DR β 1 and HLA-DP β 1 molecules in the ACA and ATA subgroups.

In total, through the five presented publications we have been able to describe 13 new genetic loci associated with susceptibility to SSc: *CD247*, *IRF8*, *ATG5*, *CSK*, *GRB10*, *NOTCH4*, *JAZF1*, *KIAA0319L*, *NFKB1*, *PSD3*, *PXK*, *SAMD9L* and *SOX5* (table 2 and figure 8) [45, 66, 92]. Of these 13 loci, six were associated with the overall SSc, four with lcSSc and two with ACA production and 1 with two different signals (one in ACA and one in ATA) (table 2 and figure 8). This again manifests the lesser statistical power in the smaller subphenotypes of SSc, but even then we have been capable of describing seven new susceptibility loci with SSc subphenotypes. This suggests the more genetically homogeneous nature of these subgroups.



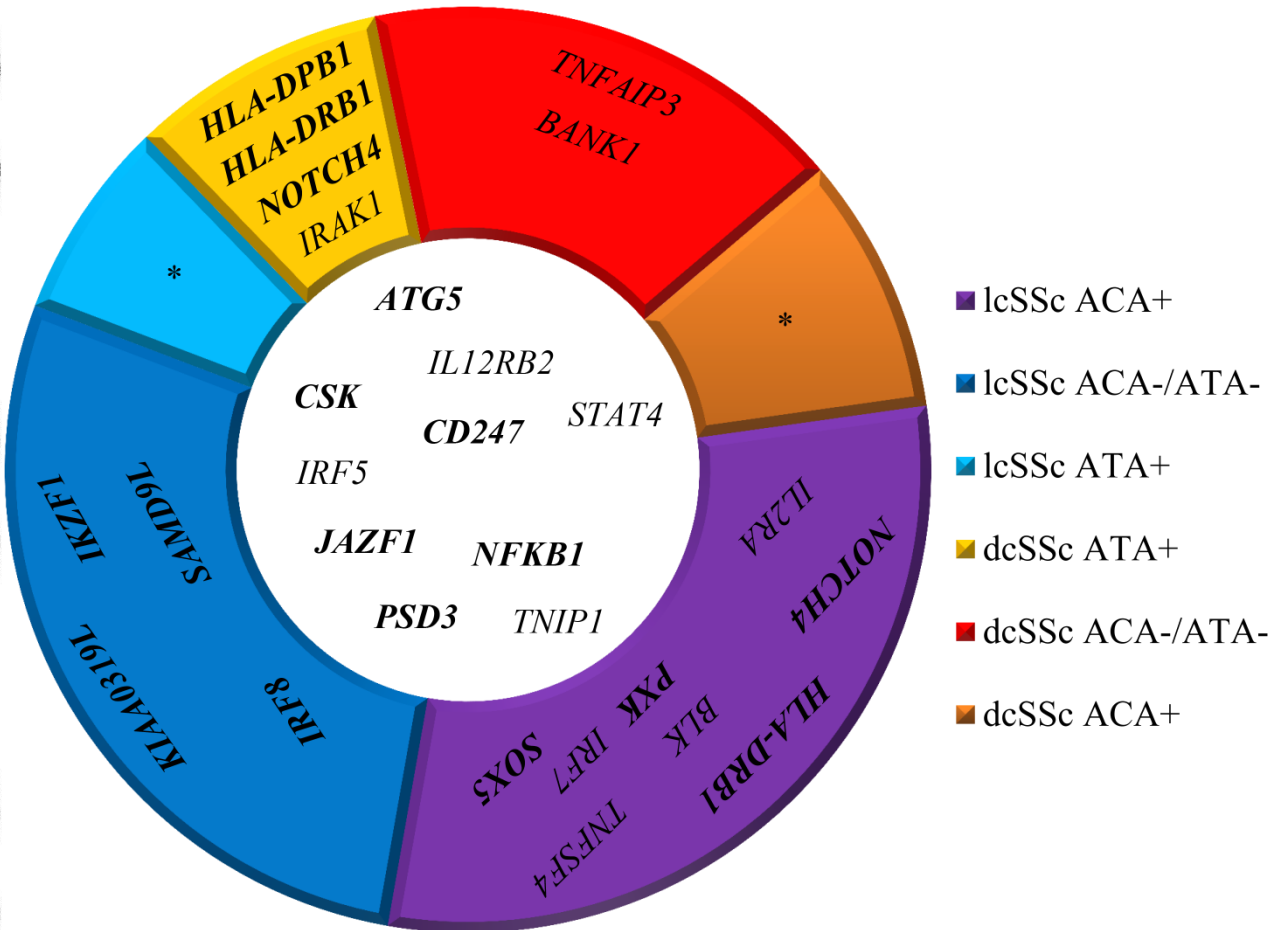
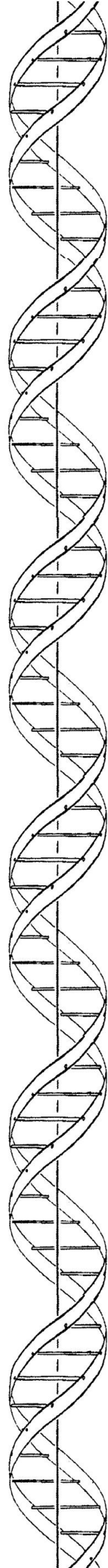


Figure 10. Diagram showing the 26 genetic loci associated with SSc or its considered subphenotypes prior to this thesis divided according disease subtype and/or auto-antibody production. *The lcSSc/ATA+ and dcSSc/ACA+ are not traditionally analyzed, because although this combinations of subtype and auto-antibody do exist, lcSSc is more commonly accompanied by ACA and dcSSc is more commonly accompanied by ATA; thus, the ACA and ATA specific associations (depicted here in the lcSSc/ACA and dcSSc/ATA) should correspond also to this segment. Marked in **bold** are the new susceptibility genetic loci described in this thesis.

Gene	Variation	Phenotype	OR	P value	References
<i>ATG5</i>	rs3827644	SSc	1.13	5.30x10 ⁻⁷	‡
<i>BANK1</i>	rs17266594	dcSSc	1.23	1.00x10 ⁻³	[47, 48]
<i>BLK</i>	rs2736340	ACA	1.47	2.20x10 ⁻⁶	[49-51]
<i>CD247</i>	rs2056626	SSc	0.82	2.09x10 ⁻⁷	[45, 91]
<i>CSK</i>	rs1378942	SSc	1.20	5.04x10 ⁻¹²	[66]
<i>HLA-DPB1</i>	AA76	ATA	8.69	3.61x10 ⁻⁷⁰	†
<i>HLA-DRB1</i>	AA58	ATA	2.82	5.60x10 ⁻³²	†
<i>HLA-DRB1</i>	AA86	ATA	1.32	1.02x10 ⁻⁴	†
<i>HLA-DRB1</i>	AA67	ATA	0.43	7.35x10 ⁻²⁵	†
<i>HLA-DRB1</i>	AA13	ACA	2.21	2.89x10 ⁻⁴⁵	†
<i>HLA-DRB1</i>	AA60	ACA	0.38	1.37x10 ⁻¹⁹	†
<i>HLA-DRB1</i>	AA71	ACA	0.70	2.02x10 ⁻¹⁰	†
<i>IKZF1</i>	rs12540874	lcSSc	1.15	1.27x10 ⁻⁶	[92]
<i>IL12RB2</i>	rs3790567	SSc	1.17	2.82x10 ⁻⁹	[56]
<i>IL2RA</i>	rs2104286	ACA	1.30	2.07x10 ⁻⁴	[57]
<i>IRAK1</i>	rs1059702	ATA	1.43	9.39x10 ⁻⁵	[58]
<i>IRF5</i>	rs10488631	SSc	1.50	1.86x10 ⁻¹³	[45, 59, 60]
<i>IRF7</i>	rs1131665	ACA	0.78	6.14x10 ⁻⁴	[61]
<i>IRF8</i>	rs11642873	lcSSc	0.75	2.32x10 ⁻¹²	[92]
<i>JAZF1</i>	rs1635852	SSc	1.09	1.11x10 ⁻⁸	‡
<i>KIAA0319L</i>	rs2275247	lcSSc	1.45	3.31x10 ⁻¹¹	‡
<i>NFKB1</i>	rs1598859	SSc	1.14	1.03x10 ⁻⁶	[66]
<i>NOTCH4</i>	rs443198	ACA	0.55	8.84x10 ⁻²¹	[92]
<i>NOTCH4</i>	rs9296015	ATA	0.54	1.14x10 ⁻⁸	[92]
<i>PSD3</i>	rs10096702	SSc	1.36	3.18x10 ⁻⁷	[66]
<i>PXK</i>	rs2176082	ACA	1.08	3.37x10 ⁻¹¹	‡
<i>SAMD9L</i>	rs1133906	lcSSc	1.07	3.17x10 ⁻⁷	‡
<i>SOX5</i>	rs11047102	ACA	1.36	1.39x10 ⁻⁷	[92]
<i>STAT4</i>	rs3821236	SSc	1.30	3.37x10 ⁻⁹	[62-64]
<i>TNFAIP3</i>	rs5029939	dcSSc	1.46	2.29x10 ⁻⁶	[65, 66]
<i>TNFSF4</i>	rs12039904	ACA	1.22	2.09x10 ⁻³	[67, 68]
<i>TNIP1</i>	rs4958881	ATA	1.19	3.26x10 ⁻⁵	[69, 70]

Table 2. All SSc associations confirmed as of the writing of this thesis. **IKZF1* association was first described by Gorlova *et al.* (one of the works presented in this thesis) for its neighbor gene *GRB10*, but the GRAIL analysis of the newly and already discovered susceptibility loci proved *IKZF1* as the most suitable character in this region. **The OR is always referring to the minor allele of the variation. †This findings have been presented in the article presented in this thesis entitled ‘Seven aminoacids in HLA-DRB1 and HLA-DPB1 explain the majority of MHC associations with systemic sclerosis’ which is still under review. ‡This findings have been presented in the article presented in this thesis entitled ‘Systemic sclerosis and systemic lupus erythematosus pan-meta-GWAS reveals six new shared susceptibility loci’ which is still under review. Marked in **bold** are the new susceptibility genetic loci described in this thesis.



Biological Relevance

Following the GO terms associated with each of the 26 genetic loci associated with as of the writing of this thesis shows that they can group in six different major categories: innate immunity, inflammation, fibrosis, B cells and T cells biology and autophagy (*figures 11 and 12*). Several of these genes have pleiotropic effects, but they have been assigned to a functional compartment according to their GO terms.

It is of interest that the gene *GRB10* is in the same region as *IKZF1*, a gene previously associated with SLE in Caucasians and Asians [93, 94]. According to the GRAIL analysis, and following the current knowledge on the function of the genes, it is more plausible that the association for SSc in this region also lies with *IKZF1*, an important

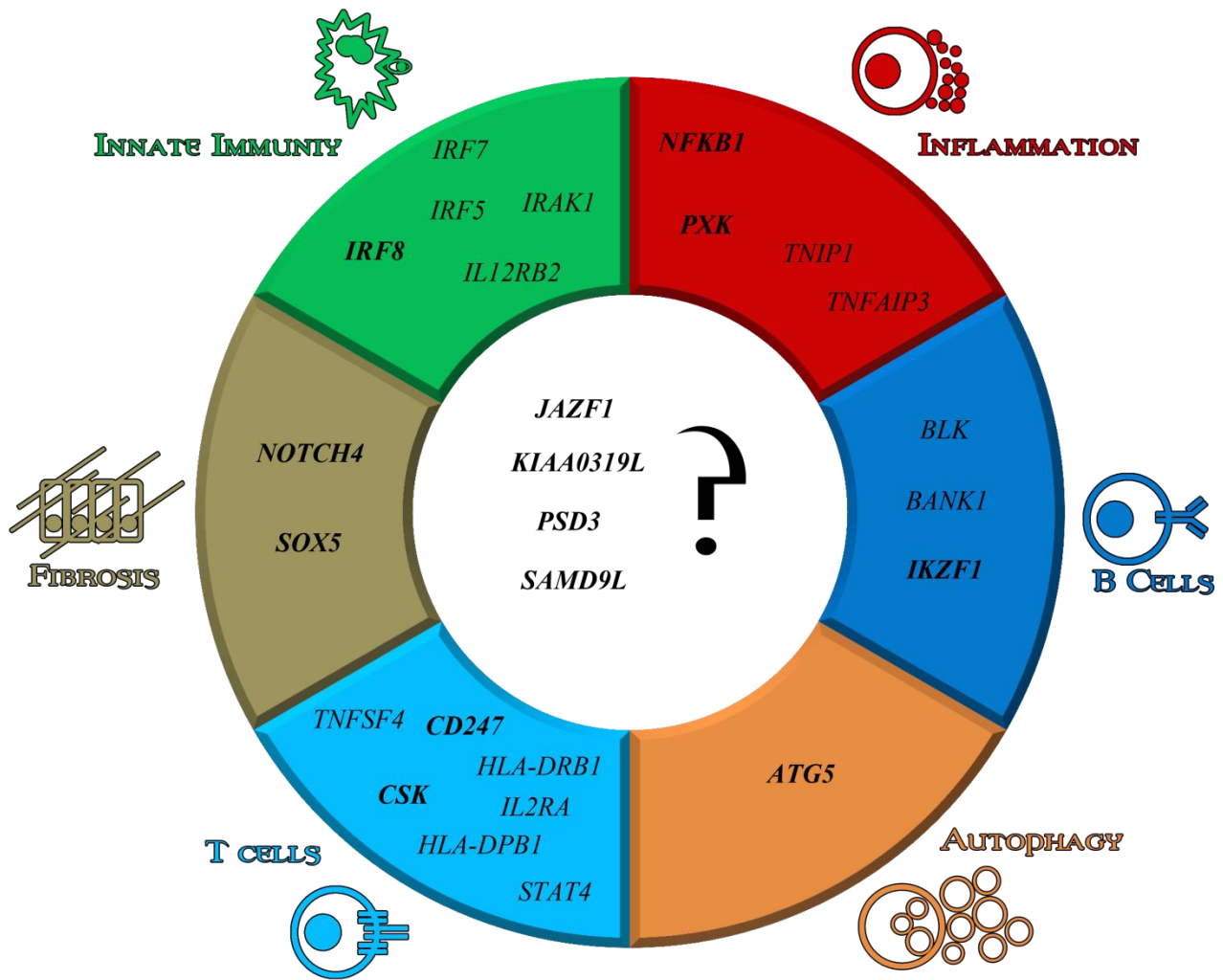


Figure 11. Distribution of the genes associated with SSc or its considered subphenotypes according to their function (based on GO terms). Marked in **bold** are the new associated loci described in this thesis.



Figure 12. Word cloud representing the GO terms of all the 26 genes associated with SSc or its considered subphenotypes as of the writing of this thesis according to GRAIL selection. The size of each GO term is weighted according to the number of occurrences in all GO terms from the 26 loci.

player in different leukemias which has confirmed roles in the biology of T and B cells [95-98], which in turn are crucial in the pathogenesis of SSc (*figure 13*). For this reason we shall address the association in this region as if *IKZF1* was the responsible gene.

Adaptive immunity plays a central role in the biology of SSc as seen in *figure 11*, with T cells, B cells and inflammation being the most represented processes to which associated genes belong (*figure 11*). Of the biological roles implied in SSc, the most well represented is that of the T cells biology, with special emphasis to antigen presentation and the corresponding signal transduction in which *HLA-DRB1*, *HLA-DPB1*, *CD247*, *CSK* and *STAT4* are directly involved (*figure 11*). Since the presence of auto-antibodies is an important feature of SSc, the presence of B cell biology among the roles of the identified genes was logical, with the participation of *BLK*, *BANK1* and *IKZF1* (*figure 11*). We also find the genes *NFKB1*, *PXK*, *TNIP1* and *TNFAIP3* as major players in SSc pathogenesis, branding inflammation as another central process. Innate immunity also plays an important role, mainly through type I interferon pathways, through the genes *IRF5*, *IRF7* and *IRF8*. The genes *IRAK1* and *IL12RB2* may act as connectors between adaptive and innate immunity in SSc pathogenesis [56, 58].

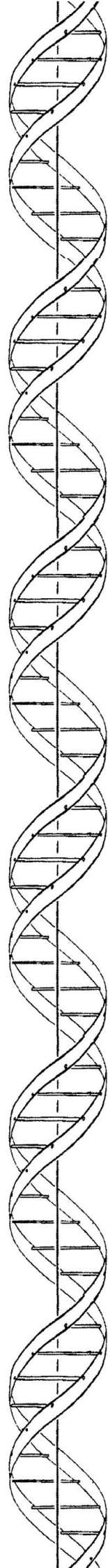
Among the newly associated genes we also find *NOTCH4* and *SOX5*, the first to susceptibility loci for SSc which are involved in the deposit of collagen [99-101], and

thus, could play role in one of the major hallmarks of SSc: fibrosis, which has been orphaned of genetic predisposition loci until now. Nonetheless, collagen deposit or fibrosis still do not appear as major player in the GO terms which represent the SSc associated genes (*figure 12*).

Among the newly discovered genes associated with either SSc or any of its considered subphenotypes we find *JAZF1*, *KIAA0319L*, *PSD3* and *SAMD9L*. The function of these genes has not been uncovered yet, and in the case of *SAMD9L*, which implication was uncovered in the pan-meta-analysis of SSc and SLE, the association is marginal in the case of SSc, pointing to a minor role in its pathogenesis. Conversely *KIAA0319L*, described in the same study, has been convincingly associated with SSc (and also SLE), not only because of its association among populations and diseases but also thanks to the fact that this gene is overexpressed in patients of both SSc and SLE compared to putatively healthy individuals. In the case of *JAZF1*, its association with SSc is supported by the fact that this gene has been already described as a genetic susceptibility loci for *SLE* [78], although its role in both diseases still remains to be uncovered. The case of *PSD3* maybe the most obscure, since its role is currently unknown and it has never been associated with any other autoimmune diseases, making this gene a suitable candidate for future studies on its function and role on the pathogenesis of SSc.

The association of the HLA class II genes alleles with many autoimmune disorders has been largely known, and many studies have been performed in SSc in order to discern such relationship [12, 52, 54, 55, 102-105]. In these studies, the main strategy has been to analyze the HLA class I and II classical alleles in case control cohorts. As reviewed in [106], the class II HLA alleles previously associated with these methods that confer genetic risk of SSc are HLA-DPB1*1301, HLA-DQB1*0501 and HLA-DRB1*1104 [10]. The only two studies of the HLA region with more than 1,000 SSc patients are those of Arnett *et al.* [12] and the one presented in this thesis, convincingly presenting the most plausible causal variation within this complex region. The study cohorts analyzed by Arnett *et al.* were composed of 1,300 cases (of which only 961 were of Caucasian ancestry) and 1,000 controls [12]. Recently, a method to impute HLA classical alleles and amino acid positions using GWAS data was developed and used to refine the association of the HLA with RA down to five amino acids in three HLA molecules [107]. We have used the same methodology to accurately impute the HLA

alleles and aminoacids in our GWAS cohorts (2,296 SSc patients and 5,356 controls), making it the largest HLA study in SSc, both in number of individuals and number of variations analyzed. With this data we have also been able to narrow down most of the association observed in the HLA region to seven aminoacids: three aminoacids in the HLA-DR β 1 molecule explain all association observed in the ACA subgroup, four aminoacids in the HLA-DP β 1 and HLA-DR β 1 molecules explain all association in the ATA subgroup, and the seven of them together explain almost all observed association with the total SSc (*table 2*).



Discerning Capability

One of the great Philosopher's stones of genetic studies on complex human traits is obtaining a set of genetic variants which can differentiate the individuals with one trait of interest from those without it with accuracy. At the point of the genetics of SSc in which we are now, with 26 confirmed genetic risk loci, we still cannot differentiate genetically patients from controls, and we may never reach that point. The main reason why we will never achieve this is because SSc (and many other classified diseases or traits) is not in any way a genetically homogenous entity. Nevertheless, when attending to more homogenous traits biologically and genetically speaking, it becomes more plausible the finding of a set of genetic variants which differentiates individuals.

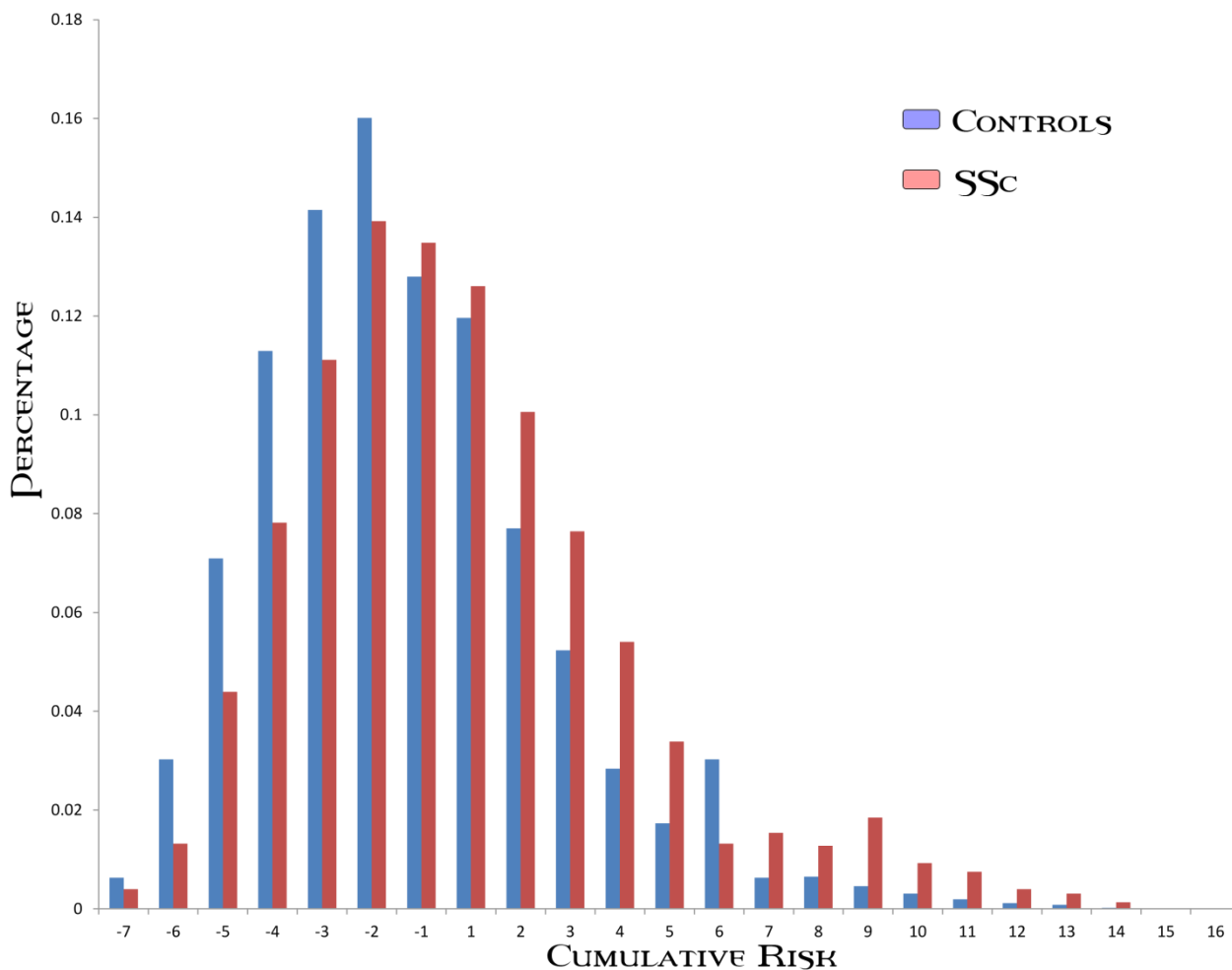


Figure 14. Bars plot of the cumulative risk presented by SSc patients and healthy controls in our GWAS cohorts. The percentage of SSc and controls are relative to the total of its own group. The genetic loci used to plot this graph were *KIAA0319L*, *IL12RB2*, *CD247*, *TNFSF4*, *STAT4*, *PXK*, *BANK1*, *NFKB1*, *TNIP1*, *NOTCH4*, *HLA-DRB1*, *HLA-DPB1*, *ATG5*, *TNFAIP3*, *JAZF1*, *IKZF1*, *SAMD9L*, *IRF5*, *BLK*, *PSD3*, *IL2RA*, *IRF7*, *SOX5*, *CSK* and *IRF8*.

To obtain a score of the risk a certain individual has of presenting any given trait, we can sum the number of susceptibility alleles known for that trait and multiply the presence of each allele for its described OR. For instance, if three polymorphic genetic loci (A/a with OR = 1.2, B/b with OR = 0.9 and C/c with OR = 1.35) are the known genetic variants that influence the trait X, we can say that an individual with the genotype aA/bB/CC has a cumulative risk of 0.1 for said trait $((1.2-1) \times 1 + (0.9-1) \times 1 + (1.35-1) \times 0 = 0.1)$. Thus, a general formula for the cumulative risk for each individual would be:

$$CR = \sum_{i=0}^n (OR_i - 1) \times a_i$$

Where CR denotes the cumulative risk, n denotes the number of known susceptibility loci for the trait, OR_i denotes the described OR for the i loci and a_i denotes the number of minor alleles the individual presents for the i loci. By subtracting 1 from the OR we accomplish that the ‘risk variants’ ($OR > 1$) increase the cumulative risk and the ‘protective variants’ ($OR < 1$) decrease the cumulative risk. When we represent in a bar plot the cumulative risk for cases and controls we obtain distributions that greatly overlap (figure 14). We can also graphically represent the distribution of the cumulative risk in all individuals by a density plot which gives us bells of distribution of cumulative risk for individuals with and without the trait (figure 15).

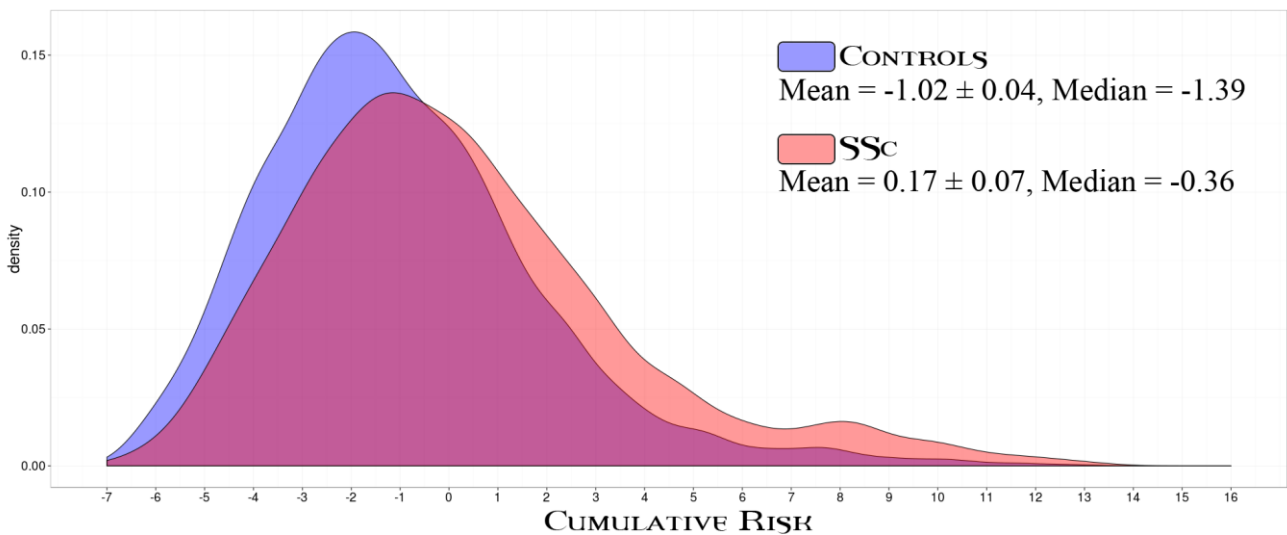
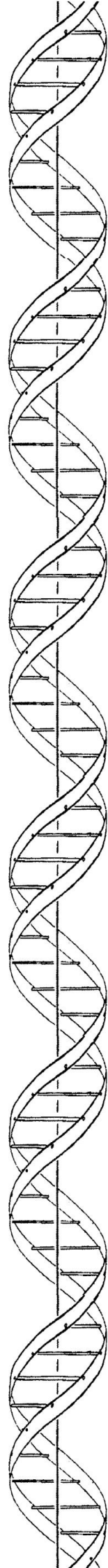


Figure 15. Density plot of the cumulative risk presented by SSc patients and healthy controls in our GWAS cohorts. The genetic loci used to calculate cumulative risk in this graph were *KIAA0319L*, *IL12RB2*, *CD247*, *TNFSF4*, *STAT4*, *PXK*, *BANK1*, *NFKB1*, *TNIP1*, *NOTCH4*, *HLA-DRB1*, *HLA-DPBI*, *ATG5*, *TNFAIP3*, *JAZF1*, *IKZF1*, *SAMD9L*, *IRF5*, *BLK*, *PSD3*, *IL2RA*, *IRF7*, *SOX5*, *CSK* and *IRF8*.

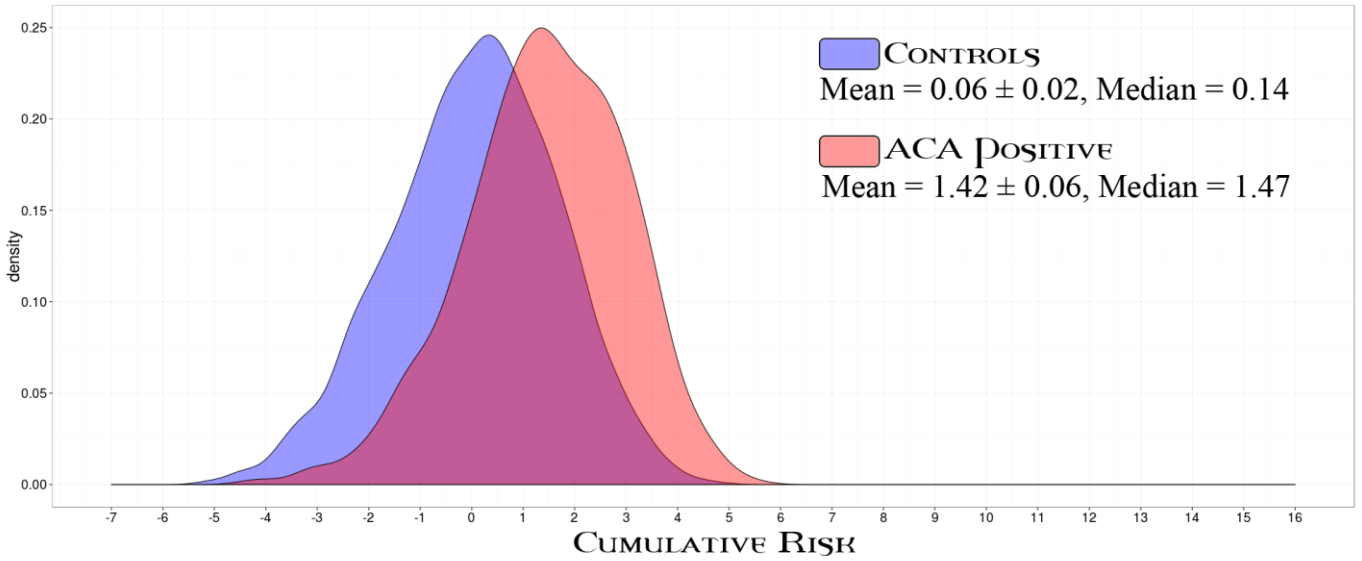
If, as previously stated, a trait is heterogeneous genetically, when we represent its described genetic variants in this fashion the differences will be minimal. This is the case that we observe in *figure 15*, where SSc patients and healthy controls are compared. Conversely, when we represent the density plot of cumulative risk for the presence of ACA or ATA we observe greater difference in the distribution of patients and controls: the difference in mean between SSc and controls is 1.19, between ACA positive and controls is 1.37 and between ATA positive and controls is 3.37 (*figure 16 and 17*). According to this, ACA and ATA production are more genetically homogeneous traits. When we go back to traits as lcSSc and dcSSc the differences in mean of the distributions fall again to 0.85 and 0.90 (*figures 16 and 17*).

Thus, as seen in *figures 16 and 17*, the very same set of genes associated either with ACA and lcSSc or ATA and dcSSc (partially overlapping groups), the discerning capability is visibly enhanced in the smaller auto-antibody positive subgroups. This difference becomes more marked when comparing the ATA positive subgroup with the dcSSc subgroup (*figure 17*).

As of now, we still cannot separate SSc patients (or any of its subphenotypes) from healthy controls by the distribution of cumulative risk. However, if the trend of discovering new susceptibility loci continues (*figure 9*), in not so many years from now, we will be able to predict the auto-antibody status of the individuals by analyzing a set of genetic markers. Furthermore, as the phenotyping of the patients of the different disorders improve, we will be able to establish the genes, pathways and set of genetic loci that identify the major hallmarks of SSc or other human disorders.



A)



B)

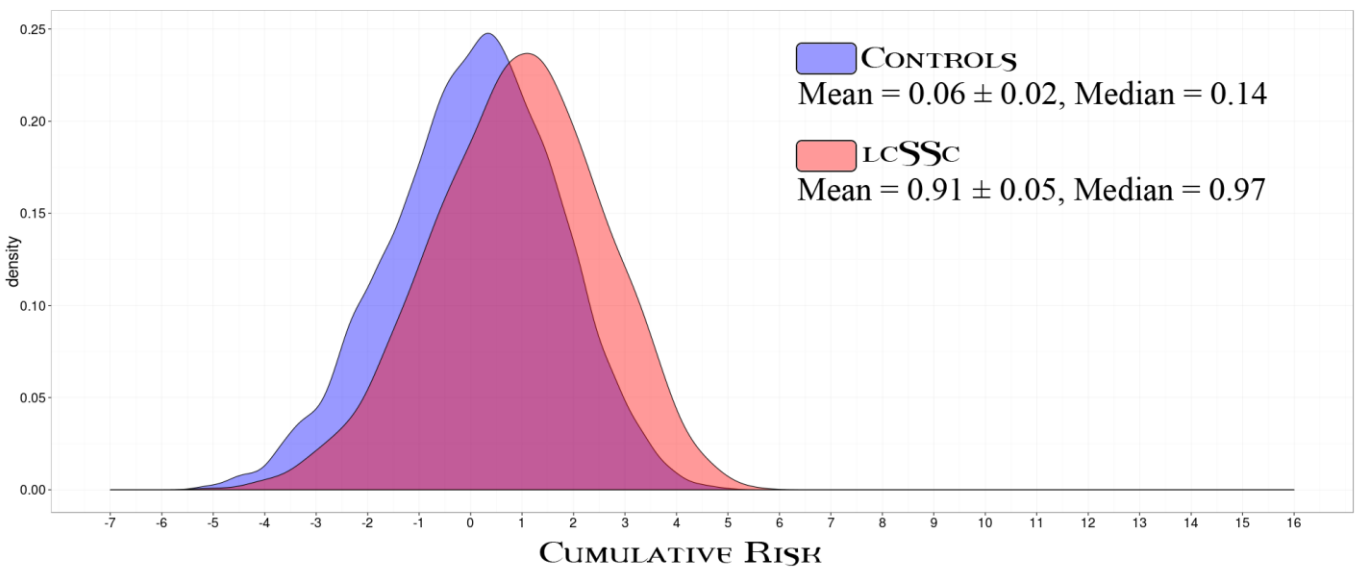


Figure 16. Density plot of the cumulative risk presented by (A) ACA positive SSc patients and controls and (B) lcSSc patients and controls in our GWAS cohorts. The set of associated genes used to calculate cumulative risk in this graph were *KIAA0319L*, *IL12RB2*, *CD247*, *TNFSF4*, *STAT4*, *PXK*, *NFKB1*, *TNIP1*, *NOTCH4*, *HLA-DRB1*, *ATG5*, *TNFAIP3*, *JAZF1*, *IKZF1*, *SAMD9L*, *IRF5*, *BLK*, *PSD3*, *IL2RA*, *IRF7*, *SOX5*, *CSK* and *IRF8*.

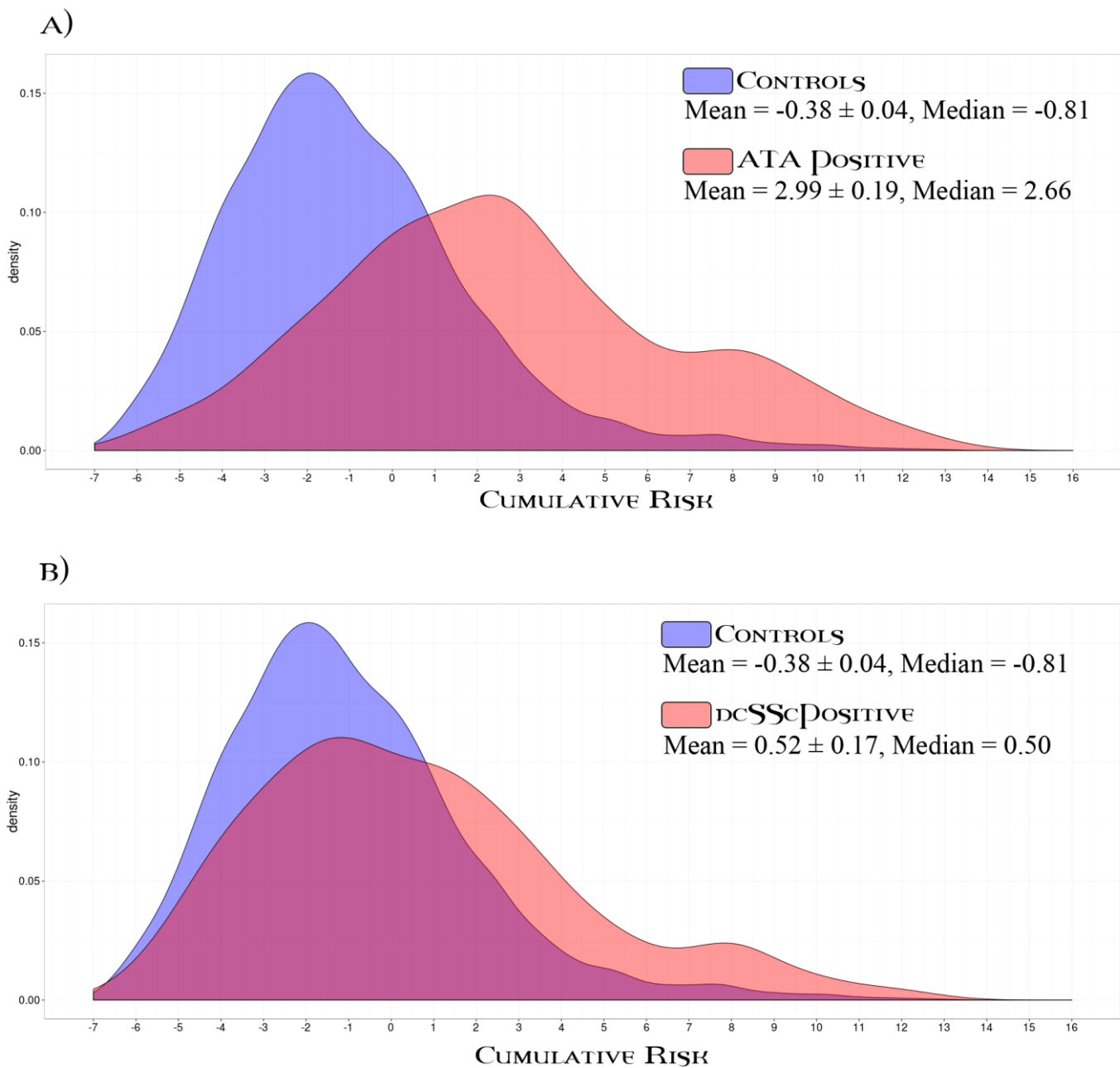


Figure 17. Density plot of the cumulative risk presented by (A) ATA positive SSc patients and controls and (B) dcSSc patients and controls in our GWAS cohorts. The set of associated genes used to calculate cumulative risk in this graph were *IL12RB2*, *CD247*, *STAT4*, *BANK1*, *NFKB1*, *TNIP1*, *NOTCH4*, *HLA-DRB1*, *HLA-DPB1*, *ATG5*, *TNFAIP3*, *JAZF1*, *IRF5*, *PSD3* and *CSK*.

PAN-AUTOIMMUNITY

Previous knowledge and part of the work exposed in this thesis show that the genetic component of SSc and SLE is greatly overlapped. When representing the OR of SSc (*table 2*) and SLE (*table 3*) in a bidimensional plot we observe that most of the confirmed genetic loci for both diseases are in common (*figure 18*).

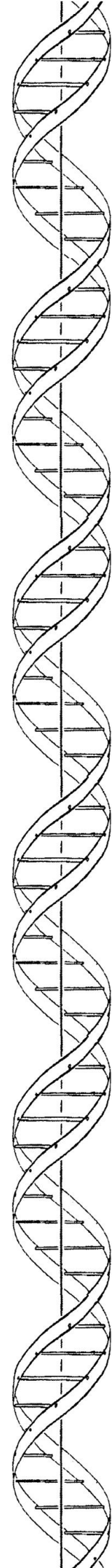
SSc and SLE have been largely known to be similar diseases in both the genetic component and the pathways involved [108]. The list of confirmed common genetic susceptibility loci extends to *ATG5*, *BANK1*, *BLK*, *CD247*, *CSK*, *HLA-DRB1*, *IKZF1*, *IL12RB2*, *IL2RA*, *IRAK1*, *IRF5*, *IRF7*, *IRF8*, *JAZF1*, *KIAA0319L*, *NOTCH4*, *PXK*, *SAMD9L*, *STAT4*, *TNFAIP3*, *TNFSF4* and *TNIP1* (*tables 2* and *3*). Nevertheless there must be differences in genetic component and pathogenic mechanisms, should not they be the same disorder also in the clinical level. In this line, it is noteworthy that the main difference in the genetic component of SSc and SLE are the most associated variations: the HLA class II alleles. Other exclusive known associations include *FCGR2A*, *ICAI1*, *IL10*, *PTTG1*, *TYK2*, *UBE2L3* and *UHRF1BP1* for SLE and *NFKB1*, *PSD3*, and *SOX5* for SSc (*tables 2* and *3*). From this, it can be deduced that the main genetic difference between SLE and SSc is what self-antigen cause the autoimmune reaction and how.

As for the pathways involved in both diseases, clues can be obtained from the word clouds representing in a weighted manner which GO terms are associated with the aforementioned exclusive susceptibility genetic loci of SSc and SLE (*figure 19*). For instance, as we can tell from the figure, specifics of the T cell biology may play a more important role in SSc. Conversely, according to the GO terms associated to SLE exclusive susceptibility loci, the cytolysis or the gamma-delta T cell activation could be more important to SLE. Thus, being involved in both diseases, the activation of T cells could be performed by different ways in the pathogenic mechanisms of SSc and SLE.

It is noteworthy that in our GWAS data the SLE exclusively confirmed loci *ICAI1*, *TYK2*, *UHRF1BP1*, *PTTG1* and *UBE2L3* have ORs > 1.1 in SSc, although they did not met the association criteria in the corresponding studies (*figure 18*). With the increasing power in SSc studies thanks to the collection of larger cohorts from different countries, it is possible that this loci will reach statistical significance in the future, thus being added to the genetic susceptibility loci shared by both disorders.

Gene	SNP	OR*	P value	Reference
<i>ATG5</i>	rs6568431	1.20	7.10x10 ⁻¹⁰	[78]
<i>BANK1</i>	rs10516487	1.11	8.30x10 ⁻⁴	[78]
<i>BLK</i>	rs2736340	1.35	7.90x10 ⁻¹⁷	[78]
<i>CD247</i>	1052231	1.20	1.03x10 ⁻²	[109]
<i>CSK</i>	rs34933034	1.32	3.35x10 ⁻⁸	[110]
<i>FCGR2A</i>	rs1801274	1.16	4.10x10 ⁻⁴	[78]
<i>IKZF1</i>	rs2366293	1.20	2.33x10 ⁻⁹	[93]
<i>HLA-DRB1</i>	0301	1.87	1.17x10 ⁻⁵⁸	[111]
<i>NOTCH4</i>	rs8192591	0.60	8.00x10 ⁻⁹	[111]
<i>MICB</i>	rs2246618	1.28	4.80x10 ⁻¹²	[111]
<i>ICA1</i>	rs10156091	1.16	6.50x10 ⁻⁴	[78]
<i>IL10</i>	rs3024505	1.19	4.00x10 ⁻⁸	[78]
<i>IL12RB2</i>	rs1874791	1.18	3.40x10 ⁻⁷	[78]
<i>IL2RA</i>	rs11594656	0.60	1.00x10 ⁻⁴	[112]
<i>IRAK1</i>	rs2269368	1.11	7.50x10 ⁻⁷	[78]
<i>IRF5</i>	rs2070197	1.88	5.80x10 ⁻²⁴	[78]
<i>IRF7</i>	rs4963128	1.20	4.90x10 ⁻⁹	[78]
<i>IRF8</i>	rs12444486	1.16	1.90x10 ⁻⁷	[78]
<i>ITGAM</i>	rs11860650	1.43	1.90x10 ⁻²⁰	[78]
<i>JAZF1</i>	rs849142	1.19	1.50x10 ⁻⁹	[78]
<i>KIAA0319L</i>	rs2275247	1.49	1.15x10 ⁻⁵	†
<i>PTPN22</i>	rs2476601	1.35	3.40x10 ⁻¹²	[78]
<i>PTTG1</i>	rs2431099	1.15	1.60x10 ⁻⁶	[78]
<i>PXK</i>	rs2176082	1.17	1.20x10 ⁻⁵	[78]
<i>SAMD9L</i>	rs1133906	1.23	1.55x10 ⁻⁵	†
<i>STAT4</i>	rs7574865	1.57	1.40x10 ⁻⁴¹	[78]
<i>TNFAIP3</i>	rs5029937	1.71	5.30x10 ⁻¹³	[78]
<i>TNFSF4</i>	rs2205960	1.22	6.30x10 ⁻⁹	[78]
<i>TNIP1</i>	rs7708392	1.27	3.80x10 ⁻¹³	[78]
<i>TYK2</i>	rs280519	1.13	7.40x10 ⁻⁵	[78]
<i>UBE2L3</i>	rs5754217	1.20	2.30x10 ⁻⁶	[78]
<i>UHRF1BP1</i>	rs11755393	1.17	2.20x10 ⁻⁸	[78]

Table 3. All SLE associations confirmed in well powered studies and cohorts from more than one country as of the writing of this thesis. *The OR is always referring to the minor allele of the variation. †This findings have been presented in the article presented in this thesis entitled ‘*Systemic sclerosis and systemic lupus erythematosus pan-meta-GWAS reveals six new shared susceptibility loci*’, which is still under review. Marked in **bold** are the new susceptibility genetic loci described in this thesis.



We can extend this reasoning to other autoimmune disorders like Crohn's disease, RA or primary biliary cirrhosis. In each of these autoimmune disorders dozens of genetic susceptibility loci have been described [84, 85, 113-115]. All these loci are partially shared among disorders as biological processes are also partially shared. In a non-exhaustive way, for example *PTPN22* is associated with SLE, RA and Crohn's disease, not associated with primary biliary cirrhosis and may play a minor role in SSc [116-120]. Another example is *STAT4*, which is associated with SSc, SLE, primary biliary cirrhosis and RA while it is not associated with Crohn's disease [62, 114, 121, 122]. The degree to which each loci causes risk in each disease is also an important factor to take into account, *e.g.*, *IRF5* is a major player in SSc and SLE genetics, while only has a minor role in RA [59, 123, 124].

To say the least, the most interesting genetic risk locus is the HLA. All of the mentioned diseases present a peak of association in the HLA region [45, 78, 80, 113-115], however, none of them share the HLA class I and II classical alleles associated. For example, HLA-DPB1*1301 is independently associated with SSc as described in this thesis and by others [12], while it is not with RA, SLE, Crohn's disease, ankylosing spondylitis or type I diabetes; HLA-DRB1*0301 is independently associated with SLE [111] and not the others and the list goes on for the other diseases. This talks about the importance of antigen presentation in each of these diseases, and how, depending on which self-antigen is presented as alien, the pathogenic mechanisms drive the course of autoimmunity in one way or another, without forgetting the influence of all specific and shared non-HLA risk loci.

As the methodologies and phenotyping improve we may observe how shared susceptibility genes among AIDs are telling us which biological processes are common in these disorders and to what extent.

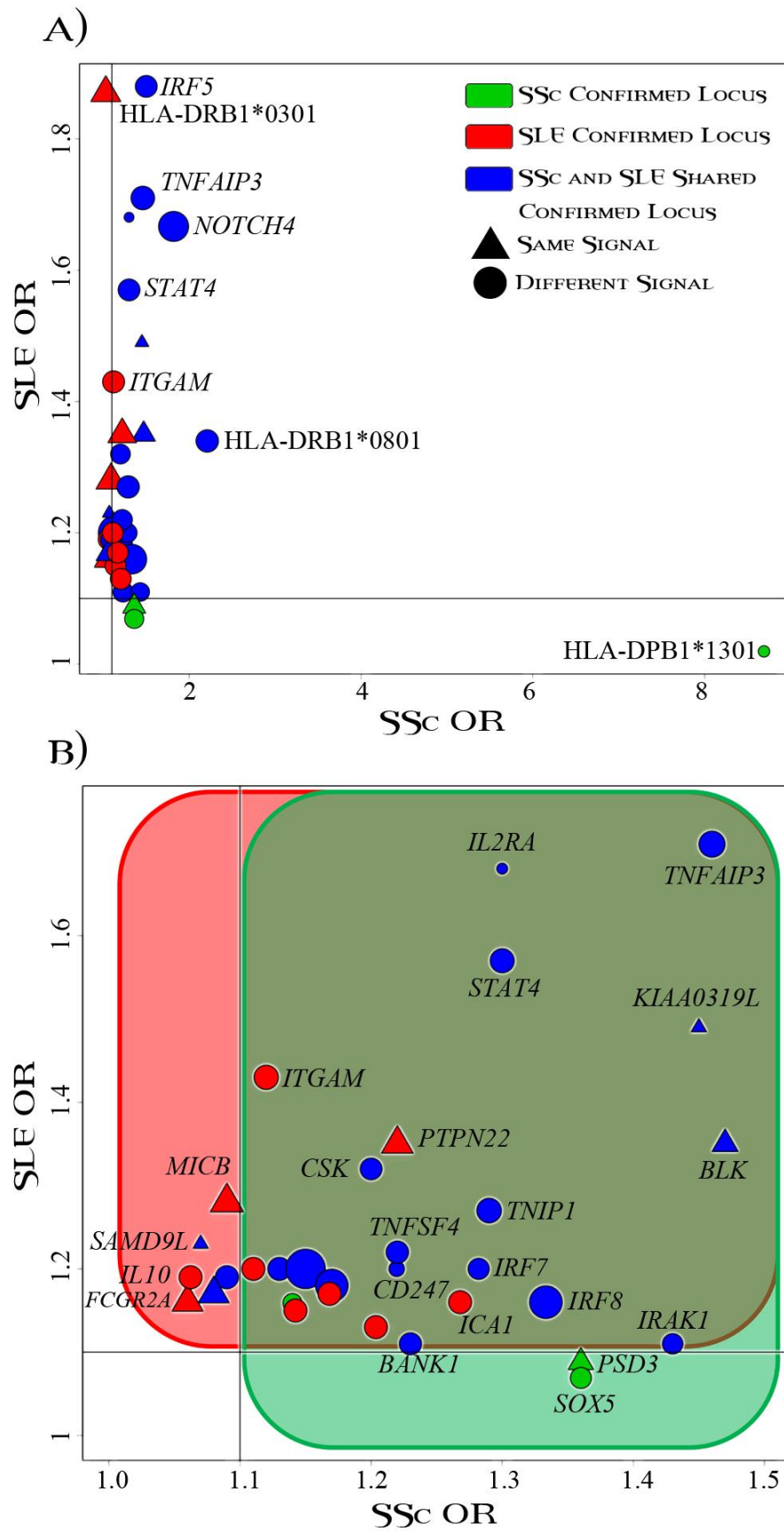
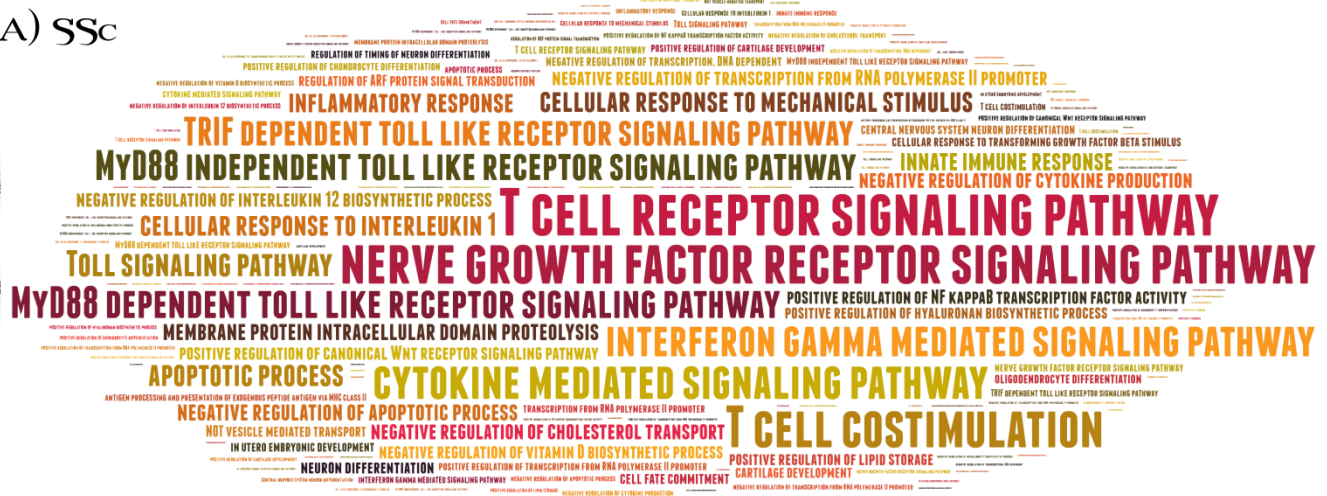


Figure 18. A) OR representation of the confirmed SSc and SLE susceptibility loci to date. B) Region of moderate associations for SSc and SLE (ORs ranging from 1 to 1.7). The size of the dots is proportional to the combined sample size of SSc and SLE.

A) SSc



B) SLE

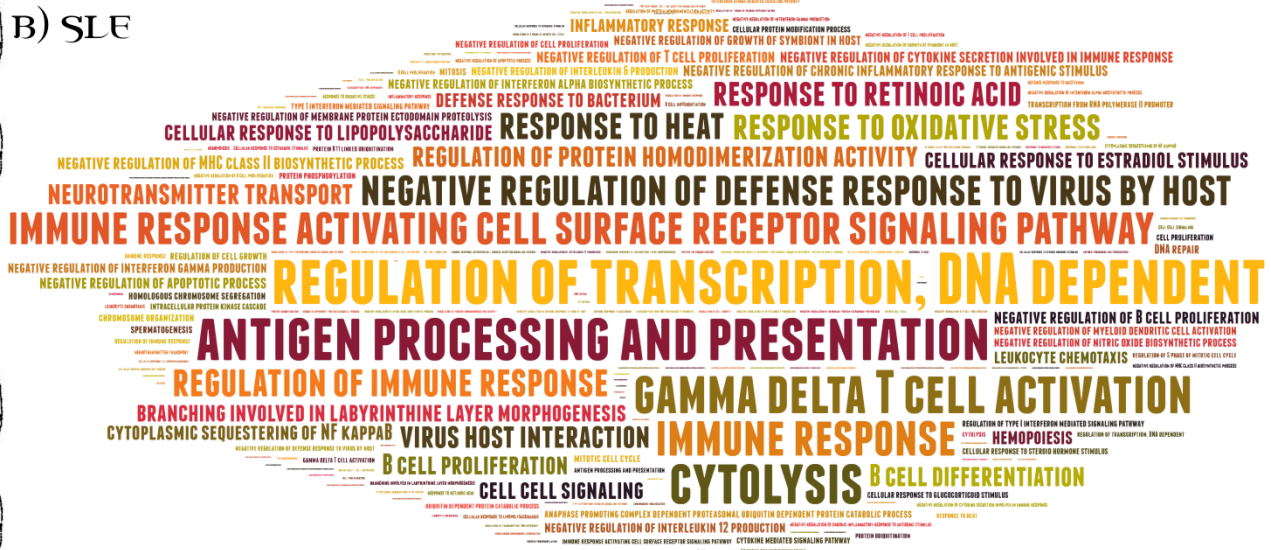


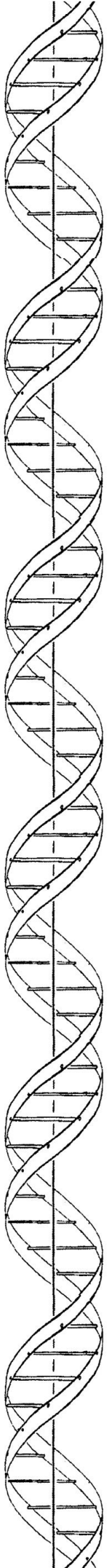
Figure 19. Word clouds representing the GO terms of A) SSc and B) SLE exclusively associated genes. The exclusively associated SSc genes included were *NFKB1*, *PSD3*, and *SOX5*. The exclusively associated SLE genes included were *FCGR2A*, *ICA1*, *IL10*, *PTTG1*, *TYK2*, *UBE2L3* and *UHRF1BP1*. The size of each GO term is weighted according to the number of occurrences in all GO terms from the implicated genes.

THE GENETIC CONTINUUM

In the genetics of complex traits, as they are studied at this point, we do not have homogeneous groups of individuals which phenotype is explained by the presence or absence of specific clinical traits, environmental factors or genetic variants. We have individuals with a unique phenotype (clinical manifestation) product of the interaction of his or her unique genetics and environment. Thus, the phenotype we encounter in each individual can overlap, according to different criteria, to a greater or lesser extent to that of another individual. As we add individuals with unique phenotypes, genotypes and environments to the equation to form what we call a disease, the genetic heterogeneity increases.

One of the many human complex traits studied to date is hair color, for which individuals can be grouped according to a certain set of rules. Regarding this phenotypic aspect we can build very simple classification criteria in which we have dark haired individuals and light haired individuals. If we perform a GWAS trying to decipher the genetic component of hair color attending to this classification, we will fail to capture all of the genetics variants influencing it but those which cause this most extreme phenotypes in the color scale, and even then we will need great sample sizes, for we will have all the genetic noise of all the ‘mid-colors’ in between included in our study. We can then do a better classification of hair color to study it, separating red, blond, brown and black hair colors.

This study has been indeed performed, and in it, several genetics variants have been identified to influence hair color, eye color and skin pigmentation [73]. These phenotypes are human complex traits influenced by many genes and environment [125]. The aforementioned study had a GWAS cohort size of 2,986 individuals genotyped with the Illumina 370k HumanHap arrays, which is far behind the sample sizes (more than 40,000 individuals) and genotyping platform (more than 1,000,000 SNPs) of the most recent RA GWASs [124]. In the study performed by Stahl et al. they find 7 new susceptibility loci for RA with ORs ranging from 1.13 to 1.29 in the GWAS level associations ($P < 5 \times 10^{-8}$). Meanwhile, in the study performed by Sulem *et al.*, they find six genetic determinants for those phenotypic traits with ORs ranging from 1.32 to 29.43 in the GWAS level associations. With a GWAS sample size of 2,986 individuals and a replication sample size of 3,932 individuals [73]. The authors conclude ‘*Our data*



on the characteristics of pigmentation are based on self-assessment, and it is likely that more complete and objective measurement techniques would strengthen the observed associations and potentially lead to further discoveries?

The color of hair we observe in individuals is consequence of the wavelength of light which is not absorbed, and thus reflected, by the pigments which are found in them. The amount, distribution and type of pigments present in hair are complex traits with its own genetic and environmental factors. And there are no single pair of individuals who have the same hair color, if we measure the color as the phenotypic continuum it truly is. A much better classification for hair color could be colorimetric measurements performed by machines under white light.

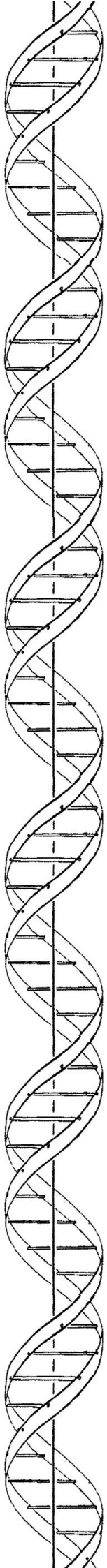
How is it that with a sample size one order of magnitude lower and an inferior genotyping technology Sulem *et al.* detected far greater effect sizes and susceptibility loci in human complex traits than the most well powered GWAS in RA? [73, 124] There are two obvious answers to this question: 1) the genetic component of RA is either more complex or weaker than that of pigmentation, or 2) the phenotype 'RA' is a far more genetically heterogeneous than the trait 'pigmentation'. Most probably the answer is a combination of both. Of course, whatever complex can be the genetics of pigmentation in humans, surely it is less complicated than the genetics of one of the two more complex systems in the human body: the immune system (together with the nervous system). Nevertheless, the genetic reality of the phenotype 'hair color' is similar as the phenotype 'autoimmune abnormality', only colors are easier to see and more intuitive to classify by our eye. We find that we cannot 'see' the true biological processes of autoimmunity as a whole like B cell auto-antibody production, auto-reactive T cell clones activation, extracellular matrix deposition and so on. But we find ourselves in the need of classify ill people in order to treat them and heal them, so we create eye observable classification criteria which are the consequence of a myriad of altered biological pathways which, in turn, are the consequence of the presence of a myriad of genetic susceptibility variants and environmental factors.

Thus, performing a GWAS in SSc is no less than trying to perform a GWAS on ethnicity. Ethnicity is the sum of a series of natural selection processes favored by the existence of reproduction barriers, which is the sum of many quantifiable biological traits, as for instance, hair color. If we compare GWAS data of Caucasians and

Blacks, we will be able to observe the genetic differences which influence the most extreme phenotypic differences between them, like skin pigmentation, but we will not observe the genetic variation which underlies in traits which are different between Caucasians and Blacks but are more continuously distributed among individuals, like height. Hence, the correct way to design the experiment is to analyze traits as homogenous as possible biologically and genetically, like skin pigment, height, or hair color. Translating this SSc, the desirable traits to analyze would be, for example, auto-antibody levels or collagen deposit (not as a binary phenotype, but the real phenotypic continuum). But in the genetics of autoimmunity we are still stuck analyzing Caucasian versus Black. This has been partially observed during the realization of this thesis, as it can be seen in *figures 15 to 17*, given the same set of susceptibility genetic loci, our ability to discern individuals with and without a trait is far greater when taking into account a biologically relevant, genetically homogenous trait as auto-antibody production (even in its binary form) than when attending to phenotypic mixture as SSc.

Thus, **individuals are the combination of many observable phenotypic continuums, which can be subdivided in many other biological continuums, which, in turn, are the product of the interaction between the genetic continuum of the involved loci and the environmental factors.** Under this theory, if we want to determine the genetic component of any given observable trait, we must divide the trait of interest in biological traits in order to determine the genetic loci for each of those, never forgetting that there is no such thing as a binary phenotype in biology and genetics.

More specifically, SSc is but a combination of observable phenotypes (Reynaud's phenomenon, digital ulcers, sclerodactyly), product of biological phenotypes (activation of autoreactive T and B cells, collagen deposit), caused by the combination of specific genetic variations (*HLA-DRB1*, *CD247*, *SOX5*) and environmental factors (silica dust, pregnancy) (*figure 21*). All these levels are continuums, not discrete traits, and this must be taken into account when delving into its genetic component. When taking a broader perspective, other traits enter into the phenotypic, biological and genetic continuums (*figure 22*).



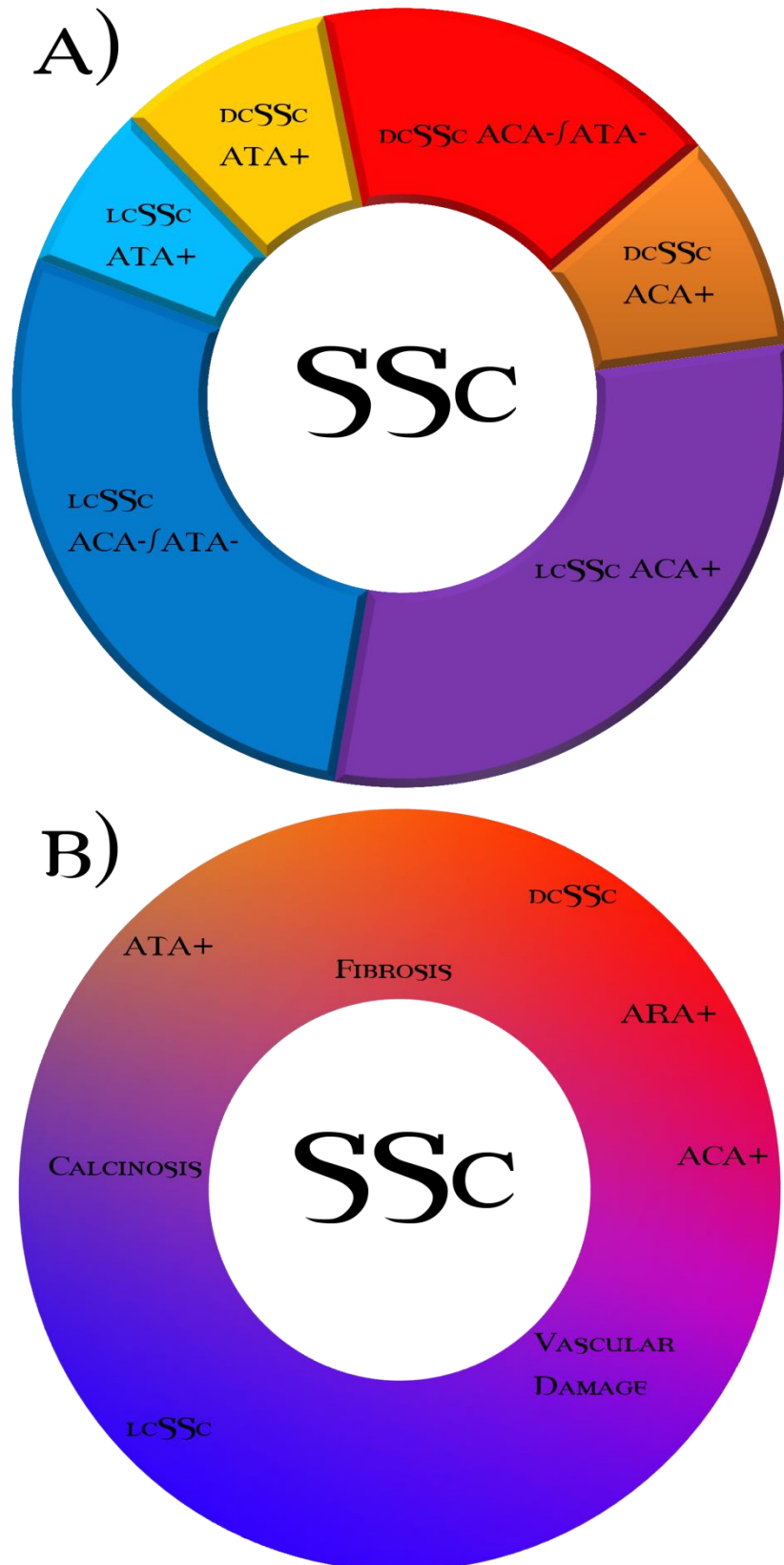


Figure 20. **A)** Classification of SSc patients subdividing them in subtypes according to presence or absence of two out of five criteria and the presence or absence at a certain fixed level of the two major auto-antibodies. **B)** Genetic continuum underlying the combination of real phenotypic continuums which are combined into what is classified as SSc.

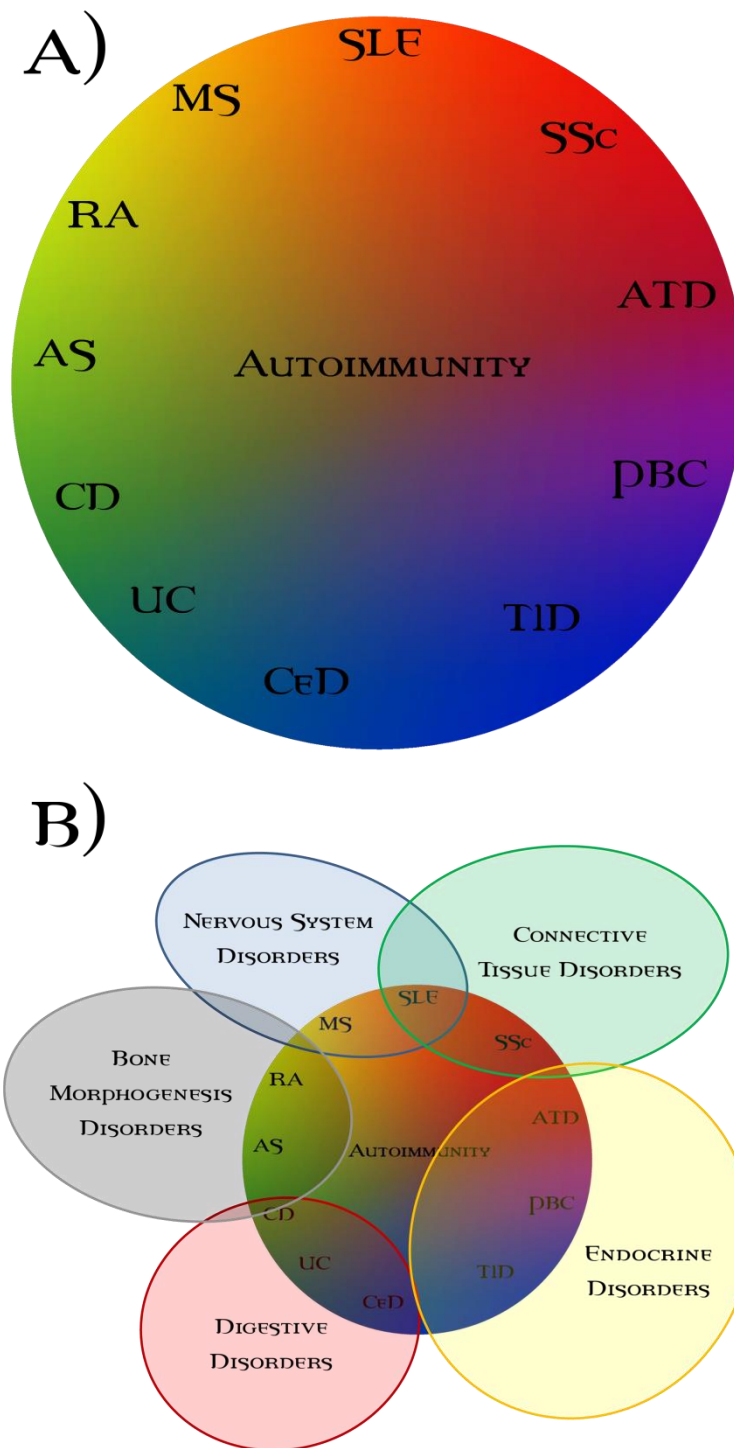


Figure 21. Two broader versions of the concept exposed in *figure 21*. **A)** All of autoimmunity is an observable phenotypic continuum, which can be subdivided into more observable phenotypic continuums (autoimmune disorders). **B)** Each of the biological continuums which compose an autoimmune disorder is shared by other abnormalities in other systems than the immune. **SLE**: systemic lupus erythematosus, **SSc**: Systemic Sclerosis, **ATD**: autoimmune thyroid disease, **PBC**: primary biliary cirrhosis, **T1D**: type 1 diabetes, **CeD**: celiac disease, **UC**: ulcerative colitis, **CD**: Crohn's disease, **AS**: ankylosing spondylitis, **RA**: rheumatoid arthritis, **MS**: multiple sclerosis

FUTURE DIRECTIONS

Five essential aspects must be improved in the field of the genetics of complex traits:

One, as the scientific community wanders in awe of the new genotyping technologies and how many complete genomes they can run so fast and with so little cost, the computers that must do the analyses of those data have become obsolete. Following the tight collaborations of pioneer scientific teams around the world collaborating back to back with biotechnology companies as Illumina, Applied Biosystems or Affimetrix, these very same bonds must be established with informatics companies such as Intel, IBM or NVidia in order to create computers for data analysis in par with the genotyping instruments themselves.

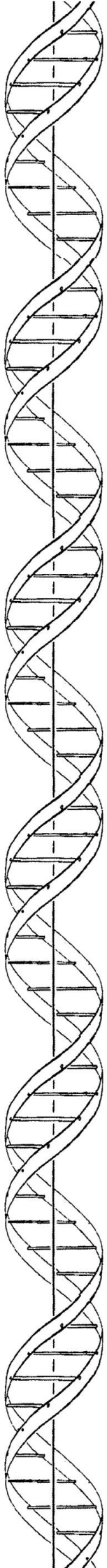
Two, the statistical power of the studies must be improved. Through inter-group collaborations around the world larger cohorts must be recruited and merged in order to genotype and analyze together as many samples as possible in order to detect the subtle genetic component of complex traits. In this line, as proven in this thesis, pan-meta-GWAS are a great tool, not without its flaws, to increase the statistical power of studies.

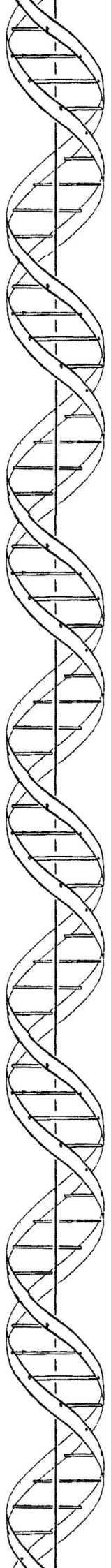
Three, in order to obtain the right answers, the right questions must be asked. There is a dire need of a better phenotyping if the genetic component of human complex traits is to be deciphered. It is of no use the analysis of 50,000 individuals who present an artificial trait with high genetic heterogeneity. Biological key processes, such as apoptosis, B cell activation or vascular damage must be phenotyped in order to collect and arrange the necessary cohorts.

Four, the increasing number of genetic susceptibility loci which are being described for complex traits must be studied in greater detail in order to discern the architecture of these associations. In this sense, the genotyping of large case/control cohorts using the custom genotyping platform ImmunoChip has already provided new insights in RA and psoriasis among others [126, 127].

Five, not only we must identify the genetic loci which confer risk to SSc and other complex traits, but we need to learn which genes in these loci and through which mechanisms they are part of the pathogenesis of this disorders. For these, the so called functional experiments must be performed to gain insight of the roles that these genes and the molecules they encode have.

In the words of the famed mathematician John Tukey, *'The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data'*.

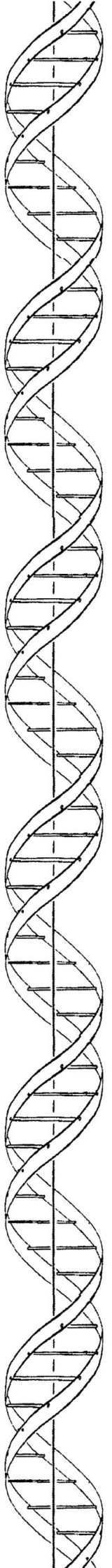




THE PURE AND SIMPLE TRUTH IS RARELY PURE
AND NEVER SIMPLE.

-OSCAR WILDE

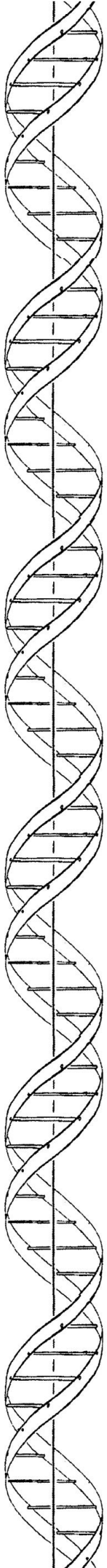
CONCLUSIONS

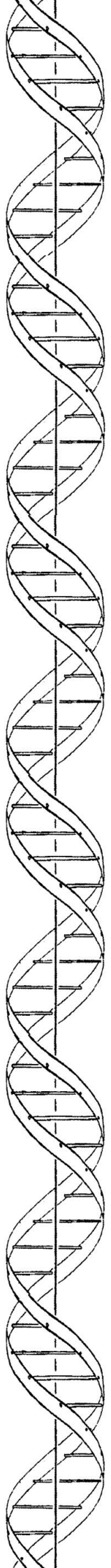


1. THE NEWLY IDENTIFIED GENES *ATG5*, *CD247*, *CSK*, *IKZF1*, *IRF8*, *JAZF1*, *KIAA0319L*, *NFKB1*, *NOTCH4*, *PSD3*, *PXK*, *SAMD9L* AND *SOX5* ARE ASSOCIATED WITH SUSCEPTIBILITY WITH SSc, ACA PRODUCTION, ATA PRODUCTION, LCSSc OR DCSSc SUSCEPTIBILITY.
2. MOST OF THE OBSERVED ASSOCIATION WITH SSc IN THE HLA REGION IS CONFINED TO THE AUTO-ANTIBODY POSITIVE SUBGROUPS AND IS EXPLAINED BY SEVEN POLYMORPHIC AMINOACIDIC POSITIONS IN THE HLA-DR β 1 AND HLA-DP β 1 MOLECULES.

3. THE **ACA** AND **ATA** PRODUCING SUBGROUPS OF **SSc** ARE MORE **GENETICALLY HOMOGENEOUS** ENTITIES THAN **SSc** AS A WHOLE, **LCSSc** OR **DCSSc**.

4. AS A GENERALIZATION OF THE PREVIOUS POINT STANDS THE UNPROVEN YET FEASIBLE THEORY THAT '**INDIVIDUALS ARE THE COMBINATION OF MANY OBSERVABLE PHENOTYPIC CONTINUUMS, WHICH CAN BE SUBDIVIDED IN MANY OTHER BIOLOGICAL CONTINUUMS, WHICH, IN TURN, ARE THE PRODUCT OF THE INTERACTION BETWEEN THE GENETIC CONTINUUM OF THE INVOLVED LOCI AND THE ENVIRONMENTAL FACTORS**'.

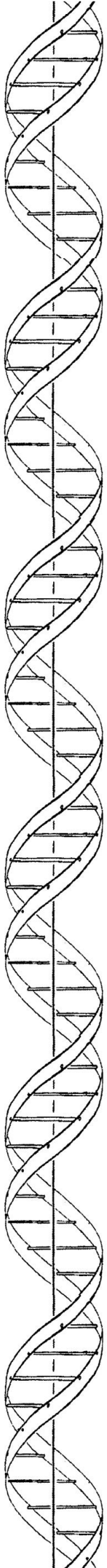




**A MIND NEEDS BOOKS AS A SWORD NEEDS A
WHETSTONE, IF IT IS TO KEEP ITS EDGE.**

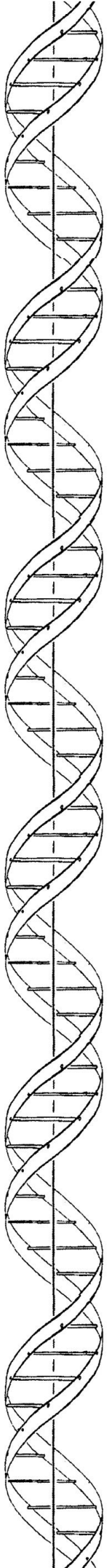
-TYRION LANNISTER

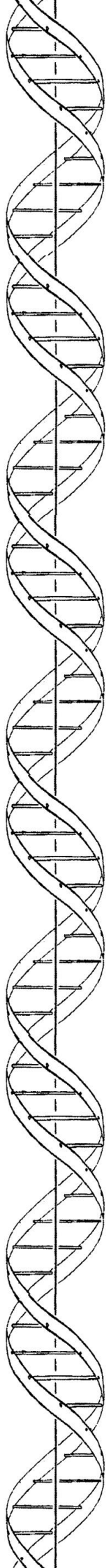
BIBLIOGRAPHY



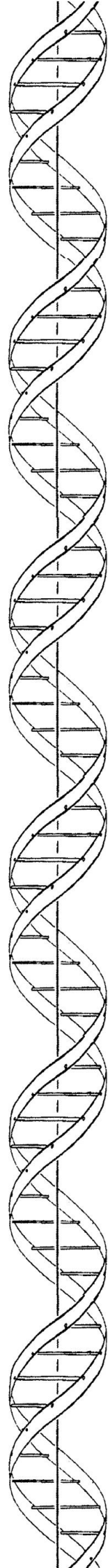
1. Gabrielli, A., E.V. Avvedimento, and T. Krieg, *Scleroderma*. N Engl J Med, 2009. **360**(19): p. 1989-2003.
2. LeRoy, E.C., et al., *Scleroderma (systemic sclerosis): classification, subsets and pathogenesis*. J Rheumatol, 1988. **15**(2): p. 202-5.
3. Prescott, R.J., et al., *Sequential dermal microvascular and perivascular changes in the development of scleroderma*. J Pathol, 1992. **166**(3): p. 255-63.
4. Fleischmajer, R. and J.S. Perlish, *Capillary alterations in scleroderma*. J Am Acad Dermatol, 1980. **2**(2): p. 161-70.
5. Harrison, N.K., et al., *Structural features of interstitial lung disease in systemic sclerosis*. Am Rev Respir Dis, 1991. **144**(3 Pt 1): p. 706-13.
6. Hoskins, L.C., et al., *Functional and morphologic alterations of the gastrointestinal tract in progressive systemic sclerosis (scleroderma)*. Am J Med, 1962. **33**: p. 459-70.
7. Perlish, J.S., G. Lemlich, and R. Fleischmajer, *Identification of collagen fibrils in scleroderma skin*. J Invest Dermatol, 1988. **90**(1): p. 48-54.
8. Fleischmajer, R., et al., *Extracellular microfibrils are increased in localized and systemic scleroderma skin*. Lab Invest, 1991. **64**(6): p. 791-8.
9. MacGregor, A.J., et al., *Pulmonary hypertension in systemic sclerosis: risk factors for progression and consequences for survival*. Rheumatology (Oxford), 2001. **40**(4): p. 453-9.
10. Martin, J.E., L. Bossini-Castillo, and J. Martin, *Unraveling the genetic component of systemic sclerosis*. Hum Genet, 2012.
11. Steen, V.D., *The many faces of scleroderma*. Rheum Dis Clin North Am, 2008. **34**(1): p. 1-15; v.
12. Arnett, F.C., et al., *Major histocompatibility complex (MHC) class II alleles, haplotypes and epitopes which confer susceptibility or protection in systemic sclerosis: analyses in 1300 Caucasian, African-American and Hispanic cases and 1000 controls*. Ann Rheum Dis, 2010. **69**(5): p. 822-7.
13. Sharif, R., et al., *Anti-fibrillarin antibody in African American patients with systemic sclerosis: immunogenetics, clinical features, and survival analysis*. J Rheumatol, 2011. **38**(8): p. 1622-30.
14. Steen, V., et al., *A clinical and serologic comparison of African-American and Caucasian patients with systemic sclerosis*. Arthritis Rheum, 2012.
15. Walker, J.G. and M.J. Fritzler, *Update on autoantibodies in systemic sclerosis*. Curr Opin Rheumatol, 2007. **19**(6): p. 580-91.
16. Voskuhl, R., *Sex differences in autoimmune diseases*. Biol Sex Differ, 2011. **2**(1): p. 1.
17. Svyryd, Y., et al., *X chromosome monosomy in primary and overlapping autoimmune diseases*. Autoimmun Rev, 2012. **11**(5): p. 301-4.
18. Ranque, B. and L. Mouthon, *Geoepidemiology of systemic sclerosis*. Autoimmun Rev, 2010. **9**(5): p. A311-8.
19. Invernizzi, P., et al., *Female predominance and X chromosome defects in autoimmune diseases*. J Autoimmun, 2009. **33**(1): p. 12-6.
20. Uz, E., et al., *Skewed X-chromosome inactivation in scleroderma*. Clin Rev Allergy Immunol, 2008. **34**(3): p. 352-5.
21. Rubtsov, A.V., et al., *Genetic and hormonal factors in female-biased autoimmunity*. Autoimmun Rev, 2010. **9**(7): p. 494-8.
22. Artlett, C.M., J.B. Smith, and S.A. Jimenez, *Identification of fetal DNA and cells in skin lesions from women with systemic sclerosis*. N Engl J Med, 1998. **338**(17): p. 1186-91.

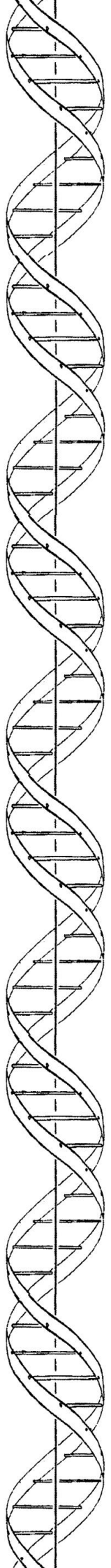
23. Nelson, J.L., et al., *Microchimerism and HLA-compatible relationships of pregnancy in scleroderma*. Lancet, 1998. **351**(9102): p. 559-62.
24. *Preliminary criteria for the classification of systemic sclerosis (scleroderma)*. Subcommittee for scleroderma criteria of the American Rheumatism Association Diagnostic and Therapeutic Criteria Committee. Arthritis Rheum, 1980. **23**(5): p. 581-90.
25. Mora, G.F., *Systemic sclerosis: environmental factors*. J Rheumatol, 2009. **36**(11): p. 2383-96.
26. McCormic, Z.D., et al., *Occupational silica exposure as a risk factor for scleroderma: a meta-analysis*. Int Arch Occup Environ Health, 2010. **83**(7): p. 763-9.
27. Ueki, A., et al., *Polyclonal human T-cell activation by silicate in vitro*. Immunology, 1994. **82**(2): p. 332-5.
28. Gourley, M. and F.W. Miller, *Mechanisms of disease: Environmental factors in the pathogenesis of rheumatic disease*. Nat Clin Pract Rheumatol, 2007. **3**(3): p. 172-80.
29. Studnicka, M.J., et al., *Pneumoconiosis and systemic sclerosis following 10 years of exposure to polyvinyl chloride dust*. Thorax, 1995. **50**(5): p. 583-5; discussion 589.
30. Juhe, S. and C.E. Lange, *[Scleroderma like skin changes. Raynaud's syndrome and acroosteolyses in workers in the polyvinyl chloride producing industry]*. Dtsch Med Wochenschr, 1972. **97**(49): p. 1922-3.
31. Yoshida, S. and M.E. Gershwin, *Autoimmunity and selected environmental factors of disease induction*. Semin Arthritis Rheum, 1993. **22**(6): p. 399-419.
32. Maul, G.G., et al., *Determination of an epitope of the diffuse systemic sclerosis marker antigen DNA topoisomerase I: sequence similarity with retroviral p30gag protein suggests a possible cause for autoimmunity in systemic sclerosis*. Proc Natl Acad Sci U S A, 1989. **86**(21): p. 8492-6.
33. Muryoi, T., et al., *Antitopoisomerase I monoclonal autoantibodies from scleroderma patients and tight skin mouse interact with similar epitopes*. J Exp Med, 1992. **175**(4): p. 1103-9.
34. *Toxic epidemic syndrome, Spain, 1981*. Toxic Epidemic Syndrome Study Group. Lancet, 1982. **2**(8300): p. 697-702.
35. Cardaba, B., et al., *Genetic approaches in the understanding of Toxic Oil Syndrome*. Toxicol Lett, 2006. **161**(1): p. 83-8.
36. Chaudhary, P., et al., *Cigarette smoking is not a risk factor for systemic sclerosis*. Arthritis Rheum, 2011. **63**(10): p. 3098-102.
37. Hudson, M., et al., *Cigarette smoking in patients with systemic sclerosis*. Arthritis Rheum, 2011. **63**(1): p. 230-8.
38. Mayes, M.D., et al., *Prevalence, incidence, survival, and disease characteristics of systemic sclerosis in a large US population*. Arthritis Rheum, 2003. **48**(8): p. 2246-55.
39. Reveille, J.D., *Ethnicity and race and systemic sclerosis: how it affects susceptibility, severity, antibody genetics, and clinical manifestations*. Curr Rheumatol Rep, 2003. **5**(2): p. 160-7.
40. Arnett, F.C., et al., *Increased prevalence of systemic sclerosis in a Native American tribe in Oklahoma. Association with an Amerindian HLA haplotype*. Arthritis Rheum, 1996. **39**(8): p. 1362-70.
41. Feghali-Bostwick, C., T.A. Medsger, Jr., and T.M. Wright, *Analysis of systemic sclerosis in twins reveals low concordance for disease and high concordance for*



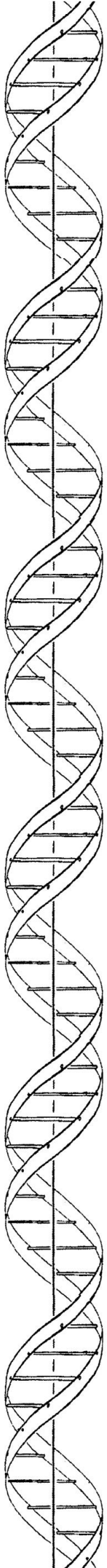
- 
- the presence of antinuclear antibodies. Arthritis Rheum, 2003. 48(7): p. 1956-63.*
42. Arnett, F.C., et al., *Familial occurrence frequencies and relative risks for systemic sclerosis (scleroderma) in three United States cohorts. Arthritis Rheum, 2001. 44(6): p. 1359-62.*
 43. Assassi, S., et al., *Clinical, immunologic, and genetic features of familial systemic sclerosis. Arthritis Rheum, 2007. 56(6): p. 2031-7.*
 44. Kuhl, P., et al., *Association of HLA antigens with progressive systemic sclerosis and morphea. Tissue Antigens, 1989. 34(3): p. 207-9.*
 45. Radstake, T.R., et al., *Genome-wide association study of systemic sclerosis identifies CD247 as a new susceptibility locus. Nat Genet, 2010. 42(5): p. 426-9.*
 46. Raychaudhuri, S., et al., *Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet, 2009. 5(6): p. e1000534.*
 47. Rueda, B., et al., *BANK1 functional variants are associated with susceptibility to diffuse systemic sclerosis in Caucasians. Ann Rheum Dis, 2010. 69(4): p. 700-5.*
 48. Dieude, P., et al., *BANK1 is a genetic risk factor for diffuse cutaneous systemic sclerosis and has additive effects with IRF5 and STAT4. Arthritis Rheum, 2009. 60(11): p. 3447-54.*
 49. Gourh, P., et al., *Association of the C8orf13-BLK region with systemic sclerosis in North-American and European populations. J Autoimmun, 2010. 34(2): p. 155-62.*
 50. Ito, I., et al., *Association of the FAM167A-BLK region with systemic sclerosis. Arthritis Rheum, 2010. 62(3): p. 890-5.*
 51. Coustet, B., et al., *C8orf13-BLK is a genetic risk locus for systemic sclerosis and has additive effects with BANK1: results from a large french cohort and meta-analysis. Arthritis Rheum, 2011. 63(7): p. 2091-6.*
 52. Simeon, C.P., et al., *Association of HLA class II genes with systemic sclerosis in Spanish patients. J Rheumatol, 2009. 36(12): p. 2733-6.*
 53. Beretta, L., et al., *Analysis of Class II human leucocyte antigens in Italian and Spanish systemic sclerosis. Rheumatology (Oxford), 2012. 51(1): p. 52-9.*
 54. Gilchrist, F.C., et al., *Class II HLA associations with autoantibodies in scleroderma: a highly significant role for HLA-DP. Genes Immun, 2001. 2(2): p. 76-81.*
 55. Kuwana, M., et al., *HLA class II genes associated with anticentromere antibody in Japanese patients with systemic sclerosis (scleroderma). Ann Rheum Dis, 1995. 54(12): p. 983-7.*
 56. Bossini-Castillo, L., et al., *A GWAS follow-up study reveals the association of the IL12RB2 gene with systemic sclerosis in Caucasian populations. Hum Mol Genet, 2012. 21(4): p. 926-33.*
 57. Martin, J.E., et al., *The autoimmune disease-associated IL2RA locus is involved in the clinical manifestations of systemic sclerosis. Genes Immun, 2012. 13(2): p. 191-6.*
 58. Dieude, P., et al., *Evidence of the contribution of the X chromosome to systemic sclerosis susceptibility: association with the functional IRAK1 196Phe/532Ser haplotype. Arthritis Rheum, 2011. 63(12): p. 3979-87.*
 59. Dieude, P., et al., *Association between the IRF5 rs2004640 functional polymorphism and systemic sclerosis: a new perspective for pulmonary fibrosis. Arthritis Rheum, 2009. 60(1): p. 225-33.*

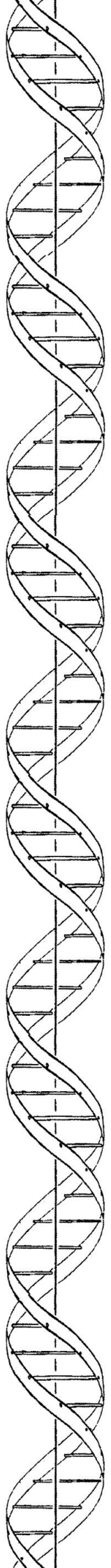
60. Ito, I., et al., *Association of a functional polymorphism in the IRF5 region with systemic sclerosis in a Japanese population*. Arthritis Rheum, 2009. **60**(6): p. 1845-50.
61. Carmona, F.D., et al., *Novel identification of the IRF7 region as an anticentromere autoantibody propensity locus in systemic sclerosis*. Ann Rheum Dis, 2012. **71**(1): p. 114-9.
62. Rueda, B., et al., *The STAT4 gene influences the genetic predisposition to systemic sclerosis phenotype*. Hum Mol Genet, 2009. **18**(11): p. 2071-7.
63. Dieude, P., et al., *STAT4 is a genetic risk factor for systemic sclerosis having additive effects with IRF5 on disease susceptibility and related pulmonary fibrosis*. Arthritis Rheum, 2009. **60**(8): p. 2472-9.
64. Tsuchiya, N., et al., *Association of STAT4 polymorphism with systemic sclerosis in a Japanese population*. Ann Rheum Dis, 2009. **68**(8): p. 1375-6.
65. Dieude, P., et al., *Association of the TNFAIP3 rs5029939 variant with systemic sclerosis in the European Caucasian population*. Ann Rheum Dis, 2010. **69**(11): p. 1958-64.
66. Martin, J.E., et al., *Identification of CSK as a systemic sclerosis genetic risk factor through Genome Wide Association Study follow-up*. Hum Mol Genet, 2012. **21**(12): p. 2825-35.
67. Bossini-Castillo, L., et al., *A replication study confirms the association of TNFSF4 (OX40L) polymorphisms with systemic sclerosis in a large European cohort*. Ann Rheum Dis, 2011. **70**(4): p. 638-41.
68. Gourh, P., et al., *Association of TNFSF4 (OX40L) polymorphisms with susceptibility to systemic sclerosis*. Ann Rheum Dis, 2010. **69**(3): p. 550-5.
69. Allanore, Y., et al., *Genome-wide scan identifies TNIP1, PSORS1C1, and RHOB as novel risk loci for systemic sclerosis*. PLoS Genet, 2011. **7**(7): p. e1002091.
70. Bossini-Castillo, L., et al., *Confirmation of TNIP1 but not RHOB and PSORS1C1 as systemic sclerosis risk factors in a large independent replication study*. Ann Rheum Dis, 2012.
71. Klein, R.J., et al., *Complement factor H polymorphism in age-related macular degeneration*. Science, 2005. **308**(5720): p. 385-9.
72. Lindgren, C.M., et al., *Genome-wide association scan meta-analysis identifies three Loci influencing adiposity and fat distribution*. PLoS Genet, 2009. **5**(6): p. e1000508.
73. Sulem, P., et al., *Genetic determinants of hair, eye and skin pigmentation in Europeans*. Nat Genet, 2007. **39**(12): p. 1443-52.
74. Weedon, M.N., et al., *A common variant of HMGA2 is associated with adult and childhood height in the general population*. Nat Genet, 2007. **39**(10): p. 1245-50.
75. *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls*. Nature, 2007. **447**(7145): p. 661-78.
76. Bezzina, C.R., et al., *Genome-wide association study identifies a susceptibility locus at 21q21 for ventricular fibrillation in acute myocardial infarction*. Nat Genet, 2010. **42**(8): p. 688-91.
77. Dunlop, M.G., et al., *Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk*. Nat Genet, 2012. **44**(7): p. 770-6.
78. Gateva, V., et al., *A large-scale replication study identifies TNIP1, PRDM1, JAZF1, UHRF1BP1 and IL10 as risk loci for systemic lupus erythematosus*. Nat Genet, 2009. **41**(11): p. 1228-33.



- 
79. Ghoussaini, M., et al., *Genome-wide association analysis identifies three new breast cancer susceptibility loci*. Nat Genet, 2012. **44**(3): p. 312-8.
 80. Gregersen, P.K., et al., *REL, encoding a member of the NF-kappaB family of transcription factors, is a newly defined risk locus for rheumatoid arthritis*. Nat Genet, 2009. **41**(7): p. 820-3.
 81. Levy, D., et al., *Genome-wide association study of blood pressure and hypertension*. Nat Genet, 2009. **41**(6): p. 677-87.
 82. Visscher, P.M., et al., *Five years of GWAS discovery*. Am J Hum Genet, 2012. **90**(1): p. 7-24.
 83. Hu, X. and M. Daly, *What have we learned from six years of GWAS in autoimmune diseases, and what is next?* Curr Opin Immunol, 2012. **24**(5): p. 571-5.
 84. Barton, A. and J. Worthington, *Genetic susceptibility to rheumatoid arthritis: an emerging picture*. Arthritis Rheum, 2009. **61**(10): p. 1441-6.
 85. Gregersen, P.K., *Susceptibility genes for rheumatoid arthritis - a rapidly expanding harvest*. Bull NYU Hosp Jt Dis, 2010. **68**(3): p. 179-82.
 86. Martin, J.E., et al., *Identification of the oxidative stress-related gene MSRA as a rheumatoid arthritis susceptibility locus by genome-wide pathway analysis*. Arthritis Rheum, 2010. **62**(11): p. 3183-90.
 87. Raychaudhuri, S., et al., *Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk*. Nat Genet, 2009. **41**(12): p. 1313-8.
 88. Kurreeman, F., et al., *Genetic basis of autoantibody positive and negative rheumatoid arthritis risk in a multi-ethnic cohort derived from electronic health records*. Am J Hum Genet, 2011. **88**(1): p. 57-69.
 89. Padyukov, L., et al., *A genome-wide association study suggests contrasting associations in ACPA-positive versus ACPA-negative rheumatoid arthritis*. Ann Rheum Dis, 2011. **70**(2): p. 259-65.
 90. Viatte, S., et al., *Genetic markers of rheumatoid arthritis susceptibility in anti-citrullinated peptide antibody negative patients*. Ann Rheum Dis, 2012. **71**(12): p. 1984-90.
 91. Dieude, P., et al., *Independent replication establishes the CD247 gene as a genetic systemic sclerosis susceptibility factor*. Ann Rheum Dis, 2011. **70**(9): p. 1695-6.
 92. Gorlova, O., et al., *Identification of novel genetic markers associated with clinical phenotypes of systemic sclerosis through a genome-wide association strategy*. PLoS Genet, 2011. **7**(7): p. e1002178.
 93. Cunninghame Graham, D.S., et al., *Association of NCF2, IKZF1, IRF8, IFIH1, and TYK2 with systemic lupus erythematosus*. PLoS Genet, 2011. **7**(10): p. e1002341.
 94. Han, J.W., et al., *Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus*. Nat Genet, 2009. **41**(11): p. 1234-7.
 95. Zhang, J., et al., *The genetic basis of early T-cell precursor acute lymphoblastic leukaemia*. Nature, 2012. **481**(7380): p. 157-63.
 96. Dail, M., et al., *Mutant Ikzf1, KrasG12D, and Notch1 cooperate in T lineage leukemogenesis and modulate responses to targeted agents*. Proc Natl Acad Sci U S A, 2010. **107**(11): p. 5106-11.
 97. Caye, A., et al., *Breakpoint-specific multiplex PCR allows the detection of IKZF1 intragenic deletions and minimal residual disease monitoring in B-cell precursor acute lymphoblastic leukemia*. Haematologica, 2012.

98. Fang, C.M., et al., *Unique contribution of IRF-5-Ikaros axis to the B-cell IgG2a response*. Genes Immun, 2012. **13**(5): p. 421-30.
99. Hardie, W.D., et al., *Genomic profile of matrix and vasculature remodeling in TGF-alpha induced pulmonary fibrosis*. Am J Respir Cell Mol Biol, 2007. **37**(3): p. 309-21.
100. Lefebvre, V., R.R. Behringer, and B. de Crombrughe, *L-Sox5, Sox6 and Sox9 control essential steps of the chondrocyte differentiation pathway*. Osteoarthritis Cartilage, 2001. **9 Suppl A**: p. S69-75.
101. Bobick, B.E., et al., *The ERK5 and ERK1/2 signaling pathways play opposing regulatory roles during chondrogenesis of adult human bone marrow-derived multipotent progenitor cells*. J Cell Physiol, 2010. **224**(1): p. 178-86.
102. Muller-Hilke, B., *HLA class II and autoimmunity: epitope selection vs differential expression*. Acta Histochem, 2009. **111**(4): p. 379-81.
103. Rands, A.L., et al., *MHC class II associations with autoantibody and T cell immune responses to the scleroderma autoantigen topoisomerase I*. J Autoimmun, 2000. **15**(4): p. 451-8.
104. Kuwana, M., et al., *An immunodominant epitope on DNA topoisomerase I is conformational in nature: heterogeneity in its recognition by systemic sclerosis sera*. Arthritis Rheum, 1999. **42**(6): p. 1179-88.
105. Kuwana, M., et al., *Association of human leukocyte antigen class II genes with autoantibody profiles, but not with disease susceptibility in Japanese patients with systemic sclerosis*. Intern Med, 1999. **38**(4): p. 336-44.
106. Martin, J.E., L. Bossini-Castillo, and J. Martin, *Unraveling the genetic component of systemic sclerosis*. Hum Genet, 2012. **131**(7): p. 1023-37.
107. Raychaudhuri, S., et al., *Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis*. Nat Genet, 2012. **44**(3): p. 291-6.
108. Assassi, S., et al., *Systemic sclerosis and lupus: points in an interferon-mediated continuum*. Arthritis Rheum, 2010. **62**(2): p. 589-98.
109. Warchol, T., et al., *The CD3Z 844 T>A polymorphism within the 3'-UTR of CD3Z confers increased risk of incidence of systemic lupus erythematosus*. Tissue Antigens, 2009. **74**(1): p. 68-72.
110. Manjarrez-Orduno, N., et al., *CSK regulatory polymorphism is associated with systemic lupus erythematosus and influences B-cell signaling and activation*. Nat Genet, 2012. **44**(11): p. 1227-30.
111. Morris, D.L., et al., *Unraveling Multiple MHC Gene Associations with Systemic Lupus Erythematosus: Model Choice Indicates a Role for HLA Alleles and Non-HLA Genes in Europeans*. Am J Hum Genet, 2012. **91**(5): p. 778-93.
112. Carr, E.J., et al., *Contrasting genetic association of IL2RA with SLE and ANCA-associated vasculitis*. BMC Med Genet, 2009. **10**: p. 22.
113. Anderson, C.A., et al., *Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47*. Nat Genet, 2011. **43**(3): p. 246-52.
114. Mells, G.F., et al., *Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis*. Nat Genet, 2011. **43**(4): p. 329-32.
115. Franke, A., et al., *Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci*. Nat Genet, 2010. **42**(12): p. 1118-25.



- 
116. Milkiewicz, P., et al., *The PTPN22 1858T variant is not associated with primary biliary cirrhosis*. Tissue Antigens, 2006. **67**(5): p. 434-7.
 117. Diaz-Gallo, L.M., et al., *Analysis of the influence of PTPN22 gene polymorphisms in systemic sclerosis*. Ann Rheum Dis, 2011. **70**(3): p. 454-62.
 118. Diaz-Gallo, L.M., et al., *Differential association of two PTPN22 coding variants with Crohn's disease and ulcerative colitis*. Inflamm Bowel Dis, 2011. **17**(11): p. 2287-94.
 119. Begovich, A.B., et al., *A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis*. Am J Hum Genet, 2004. **75**(2): p. 330-7.
 120. Kyogoku, C., et al., *Genetic association of the R620W polymorphism of protein tyrosine phosphatase PTPN22 with human SLE*. Am J Hum Genet, 2004. **75**(3): p. 504-7.
 121. Diaz-Gallo, L.M., et al., *STAT4 gene influences genetic predisposition to ulcerative colitis but not Crohn's disease in the Spanish population: a replication study*. Hum Immunol, 2010. **71**(5): p. 515-9.
 122. Remmers, E.F., et al., *STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus*. N Engl J Med, 2007. **357**(10): p. 977-86.
 123. Graham, R.R., et al., *A common haplotype of interferon regulatory factor 5 (IRF5) regulates splicing and expression and is associated with increased risk of systemic lupus erythematosus*. Nat Genet, 2006. **38**(5): p. 550-5.
 124. Stahl, E.A., et al., *Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci*. Nat Genet, 2010. **42**(6): p. 508-14.
 125. Rees, J.L., *Genetics of hair and skin color*. Annu Rev Genet, 2003. **37**: p. 67-90.
 126. Eyre, S., et al., *High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis*. Nat Genet, 2012. **44**(12): p. 1336-40.
 127. Tsoi, L.C., et al., *Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity*. Nat Genet, 2012. **44**(12): p. 1341-8.

