CSIC
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

ipbln

UNIVERSIDAD
DE GRANADA

Deciphering the genomic architecture of systemic sclerosis

Doctoral Thesis

David González Serna, 2021

Programa de Doctorado en Biomedicina

Tesis Doctoral

Deciphering the genomic architecture of systemic sclerosis

Memoria presentada por el graduado en Biomedicina Básica y Experimental David González Serna para optar al grado de Doctor Internacional por la Universidad de Granada.

Programa de Doctorado en Biomedicina.

Directores:

Javier Martín Ibáñez, Profesor de Investigación del CSIC.

Ana María Márquez Ortiz, Investigadora Miguel Servet del Instituto de Investigación Biosanitaria de Granada.

CSIC
Consejo Superior de Investigaciones Científicas

Universitas Granatensis 1531
Universidad de Granada

Instituto de Parasitología y Biomedicina López-Neyra,
CSIC.
Granada, octubre de 2021

Editor: Universidad de Granada. Tesis Doctorales
Autor: David González Serna
ISBN: 978-84-1117-204-2
URI: http://hdl.handle.net/10481/72363

MINISTERIO DE CIENCIA E INNOVACIÓN

CSIC

ipbln

INSTITUTO DE PARASITOLOGÍA Y BIOMEDICINA "LÓPEZ-NEYRA"

El doctorando / *The doctoral candiate* **Davíd González Serna** y los directores de la tesis / *and the thesis supervisors*: **Javier Martín Ibáñez** y **Ana María Márquez Ortiz**

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.

/

Guarantee, by signing this doctoral thesis, that the work has been done by the doctoral candidate under the direction of the thesis supervisors and, as far as our knowledge reaches, in the performance of the work, the rights of other authors to be cited (when their results or publications have been used) have been respected

Lugar y fecha / *Place and date*:

Granada, 21 de octubre de 2021

Directores de la Tesis / *Thesis supervisors;* Doctorando / *Doctoral candidate:*

Firma / *Signed* Firma / *Signed*

A mi familia,

A mis amigos,

A la música que me acompaña

INDEX

ABBREVIATIONS 8

SUMMARY 14

INTRODUCTION 19

1. Systemic sclerosis: the complexity of autoimmunity 19

2. Environmental component of systemic sclerosis 26

2.1. Chemical agents 27

2.2. Biological agents 29

2.3. Physical agents 30

3. The multifactorial genetic component of systemic sclerosis 30

3.1. Genotyping studies 31

3.2. Transcriptomic studies 43

3.3. Epigenomic studies 46

OBJECTIVES 52

MATERIAL AND METHODS, RESULTS, AND DISCUSSION 53

- **Chapter 1: Analysis of the genetic component of systemic sclerosis in Iranian and Turkish populations through a genome-wide association study** 53

1.1. Material and methods 54

1.1.1. Study population 54

1.1.2. Genotyping 54

1.1.3. Quality control 55

1.1.4. Imputation of GWAS data 56

1.1.5.Human leukocyte antigen (HLA) imputation 56

1.1.6. Statistical analysis 57

1.2. Results 58

1.2.1. Meta-analysis 62

1.2.2. HLA analysis 62

1.3. Discussion 69

- **Chapter 2: Analysis of the shared genetic component between systemic sclerosis and Crohn's disease through a cross-disease meta-GWAS** 73

  2.1. Material and methods 74

  2.1.1. Study population 74

  2.1.2. Quality control and imputation 75

  2.1.3. Statistical analysis 76

  2.1.4. Functional annotation 78

  2.1.5. Protein-protein interaction and gene set enrichment analysis 78

  2.2. Results 79

  2.2.1. Meta-analysis and replication 79

  2.2.2. Functional effect on gene expression 83

  2.2.3. Protein-protein interaction and enrichment analysis 83

  2.3. Discussion 86

  2.4. Supplementary data 90

- **Chapter 3: Genome-wide meta-analysis study of the shared genetic component in systemic seropositive rheumatic diseases** 94

  3.1. Material and methods 95

  3.1.1. Study population 95

  3.1.2. Quality control and imputation 96

  3.1.3. Statistical analysis 96

  3.2. Results 99

  3.2.1. Cross-disease meta-analysis and disease contribution 99

  3.2.2. Associated loci and their functional enrichment on regulatory elements 104

  3.2.3. Expression quantitative trait loci (eQTL) and associated variants 106

  3.2.4. Drug target enrichment analysis 107

  3.3. Discussion 110

  3.4. Supplementary data 116

- **Chapter 4: Expression quantitative trait locus (eQTL) analysis in systemic sclerosis identifies new candidate genes associated with multiple aspects of disease pathology** 121

  4.1. Material and methods 122

4.1.1. Study population 122

4.1.2. RNA sequencing 123

4.1.3. Genotyping 124

4.1.4. eQTL detection 124

4.1.5. Analysis of SSc patients without immunomodulating medication 126

4.1.6. Stepwise linear regression and comparison of r-square values 127

4.1.7. Differential expression analysis 127

4.1.8. Transcription factor binding site analysis 128

4.1.9. Drug target analysis 129

4.1.10. Tissue enrichment analysis 129

4.1.11. Pathway enrichment analysis 130

4.2. Results 130

4.2.1. Study design, gene and eQTL numbers, and comparison to external datasets 130

4.2.2. SSc specific eQTLs 132

4.2.3. Enrichment of blood eQTLs in tissue affected by disease 134

4.2.4. Expression variance explained (EVE) can be used to prioritize SSc eQTLs and SSc eGenes 135

4.2.5. SSc eGenes grouped by the hallmarks of SSc pathogenesis 137

4.2.6. Differential expression of SSc eGenes in disease-affected tissues 138

4.2.7. Results of transcription factor binding site analysis 141

4.2.8. Drug repurposing 143

4.3. Discussion 145

- **Chapter 5: Functional genomics in primary T cells and monocytes: linking genetic susceptibility loci with mechanisms influencing systemic sclerosis risk** 151

5.1. Material and methods 152

5.1.1. Isolation of CD4$^+$ T cells and CD14$^+$ monocytes 152

5.1.2. Promoter capture Hi-C probe design 152

5.1.3. Capture Hi-C library generation 153

5.1.4. Promoter capture Hi-C (pCHi-C) sequencing and processing 154

5.1.5. RNA-seq library generation 155

5.1.6. RNA-seq data processing 156

5.1.7. Linking differential expression and differential interactions in CD4$^+$ T cells vs CD14$^+$ monocytes 157

5.1.8. Defining SSc-associated GWAS loci 157

5.1.9. Defining enhancers and TADs in CD4$^+$ T cells and CD14$^+$ monocytes ..158

5.1.10. Overlap among pCHi-C, SSc-associated GWAS loci, and enhancer regions 158

5.1.11. Visualization tools 159

5.1.12. Drug target analysis 159

5.2. Results 160

5.2.1. Differential interactions and expression between SSc patients and healthy controls 160

5.2.2. Linking differential expression and differential interactions in CD4$^+$ T cells vs CD14$^+$ monocytes 162

5.2.3. SSc-associated loci and CD4$^+$ / CD14$^+$ promoter interactions 164

5.2.4. Drug repurposing in SSc 173

5.3. Discussion 176

5.4. Supplementary data 181

GENERAL DISCUSSION 201

CONCLUSIONS 207

BIBLIOGRAPHY 212

ABBREVIATIONS

3C: Chromosome conformation capture

ACA: Anticentromere antibody

AD: Autoimmune disease

ANA: Antinuclear antibody

ARA: Anti-RNA polymerase III antibody

ATA: Anti-topoisomerase 1 antibody

BAFF: B cell activating factor

BP: Base pair

CD: Crohn's disease

ChIP: Chromatin immunoprecipitation

CI: Confidence interval

CMV: Cytomegalovirus

CPM: Counts per million

CREST: Calcinosis, Raynaud's phenomenon, esophageal dysmotility, sclerodactyly, and telangiectasia

CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats

dcSSc: Diffuse cutaneous systemic sclerosis

DHS: DNase hypersensitivity site

EBV: Epstein-Barr virus

EC: Endothelial cell

ECM: Extracellular matrix

EN-1: Endothelin-1

eQTL: Expression quantitative trait locus

EVE: Expression variance explained

FDR: False Discovery Rate

FE: Fold enrichment

FGF-2: Fibroblast growth factor-2

GO: Gene ontology

GTEx: Genotype-Tissue Expression

GWAS: Genome-wide association study

Hi-C: High-throughput chromosome conformation capture

HLA: Human leukocyte antigen

HRC: Haplotype Reference Consortium

HWE: Hardy-Weinberg equilibrium

IC: Information content

IFN: Interferon

Ig: Immunoglobulin

IIM: Idiopathic inflammatory myopathy

IL: Interleukin

ILD: Interstitial lung disease

IMID: Immune-mediated inflammatory disease

IPF: Idiopathic pulmonary fibrosis

IRF: Interferon regulatory factor

lcSSc: Limited cutaneous systemic sclerosis

LD: Linkage disequilibrium

MAF: Minor allele frequency

MIS: Michigan Imputation Server

NGS: Next-generation sequencing

OR: Odds ratio

PAH: Pulmonary arterial hypertension

PBMC: Peripheral blood mononuclear cell

PC: Principal component

PCA: Principal component analysis

pCHi-C: Promoter capture Hi-C

PDGF: Platelet-derived growth factor

PF: Pulmonary fibrosis

PIR: Promoter interacting region

PPI: Protein-protein interaction

PWM: Position weight matrices

QC: Quality control

RA: Rheumatoid arthritis

RIN: RNA integrity number

scRNA-seq: Single-cell RNA sequencing

SD: Standard deviation

SLE: Systemic lupus erythematosus

SNP: Single nucleotide polymorphism

SRC: Scleroderma renal crisis

SSc: Systemic sclerosis or scleroderma

TAD: Topologically associated domain

TALE: Transcription activator-like effector

TCR: T cell receptor

TFBS: Transcription factor binding site

TGF-β: Transforming growth factor β

Th1: T helper 1

TNF-α: Tumor necrosis factor α

Treg: Regulatory T cell

TSS: Transcription start site

VEGF: Vascular endothelial growth factor

VEP: Variant effect predictor

WES: Whole exome sequencing

WGS: Whole genome sequencing

WTCCC: Wellcome Trust Case Control Consortium

α-SMA: α-smooth muscle actin-positive

GENES

AIRE: Autoimmune Regulator

ARCN1: Archain 1

ARL14: ADP Ribosylation Factor Like GTPase 14

ATG5: Autophagy Related 5

BAK1: BCL2 Antagonist/Killer 1

BCL11B: BAF Chromatin Remodeling Complex Subunit BCL11B

BLK: BLK Proto-Oncogene, Src Family Tyrosine Kinase

C4A: Complement C4A

C4B: Complement C4B

CHD7: Chromodomain Helicase DNA Binding Protein 7

CREG1: Cellular Repressor of E1A Stimulated Genes 1

CSK: C-Terminal Src Kinase

CXCL8: C-X-C Motif Chemokine Ligand 8

CXCR5: C-X-C Motif Chemokine Receptor 5

DDX6: DEAD-Box Helicase 6

DGKQ: Diacylglycerol Kinase Theta

DNASE1L3: Deoxyribonuclease 1 Like 3

DRD4: Dopamine Receptor D4

DXO: Decapping Exoribonuclease

ELF1: E74 Like ETS Transcription Factor 1

ERAP1: Endoplasmic Reticulum Aminopeptidase 1

FAM167A: Family with Sequence Similarity 167 Member A

FGFRL1: Fibroblast growth factor receptor like 1

FLNB: Filamin B

FOXP3: Forkhead Box P3

GOT1: Glutamic-Oxaloacetic Transaminase 1

GSDMA: Gasdermin A

HCG11: HLA Complex Group 11

HIBCH: 3-Hydroxyisobutyryl-CoA Hydrolase

HID1: HID1 Domain Containing

IER3: Immediate Early Response 3

IFIH1: Interferon Induced with Helicase C Domain 1

IFT46: Intraflagellar Transport 46

IKZF3: IKAROS Family Zinc Finger 3

IL12A: Interleukin 12A

IL12RB1: Interleukin 12 Receptor Subunit Beta 1

IL12RB2: Interleukin 12 Receptor Subunit Beta 2

IL23R: Interleukin 23 Receptor

IRF1: Interferon Regulatory Factor 1

IRF5: Interferon Regulatory Factor 5

IRF7: Interferon Regulatory Factor 7

IRF8: Interferon Regulatory Factor 8

KPNA4: Karyopherin Subunit Alpha 4

LEF1: Lymphoid Enhancer Binding Factor 1

LIMK1: LIM Domain Kinase 1

MHC: Major Histocompatibility Complex

NAB1: NGFI-A Binding Protein 1

NABP1: Nucleic Acid Binding Protein 1

NCR3: Natural Cytotoxicity Triggering Receptor 3

NEU1: Neuraminidase 1

NFKB1: Nuclear Factor Kappa B Subunit 1

NKX2.3: NK2 Homeobox 3

NOTCH1: Notch Receptor 1

NOTCH4: Notch Receptor 4

ORMDL3: ORMDL Sphingolipid Biosynthesis Regulator 3

PADI2: Peptidyl Arginine Deiminase 2

PPARG: Peroxisome Proliferator Activated Receptor Gamma

PRDM1: PR/SET Domain 1

PRR12: Proline Rich 12

PSORS1C1: Psoriasis Susceptibility 1 Candidate 1

PTPN11: Protein Tyrosine Phosphatase Non-Receptor Type 11

PTPN22: Protein Tyrosine Phosphatase Non-Receptor Type 22

RAF1: Raf-1 Proto-Oncogene, Serine/Threonine Kinase

RHOB: Ras Homolog Family Member B

RUNX1: RUNX Family Transcription Factor 1

S1PR3: Sphingosine-1-Phosphate Receptor 3

SLC22A5: Solute Carrier Family 22 Member 5

SMC4: Structural Maintenance of Chromosomes 4

STAT1: Signal Transducer and Activator of Transcription 1

STAT3: Signal Transducer and Activator of Transcription 3

STAT4: Signal Transducer and Activator of Transcription 4

TAMM41: TAM41 Mitochondrial Translocator Assembly and Maintenance Homolog

TAPBP: TAP Binding Protein

TNFAIP3: TNF Alpha Induced Protein 3

TNFSF4: TNF Superfamily Member 4

TNIP1: TNFAIP3 Interacting Protein 1

TNPO3: Transportin 3

TNXB: Tenascin XB

TREX1: Three Prime Repair Exonuclease 1

TUBB: Tubulin Beta Class I

TYK2: Tyrosine Kinase 2

UBE2L3: Ubiquitin Conjugating Enzyme E2 L3

UPK2: Uroplakin 2

YDJC: YdjC Chitooligosaccharide Deacetylase Homolog

ZBTB9: Zinc Finger and BTB Domain Containing

SUMMARY

Systemic sclerosis (SSc) is a complex rheumatic autoimmune disease (AD) with an important genetic and environmental component, characterized by the triad of pathological hallmarks: extensive fibrosis of skin and internal organs, vascular damage, and altered immune response (including the presence of autoantibodies). SSc shows a wide range of phenotypical manifestations and heterogeneous clinical characteristics, making it difficult to treat.

The present PhD dissertation is focused on the study of the genetic component of SSc. To date, more than 25 loci have been firmly associated to SSc by genome-wide association studies (GWAS). Nevertheless, most of these studies have been performed in Caucasian population. In order to extent the knowledge of the genetic component of SSc, we performed a GWAS in the Iranian and Turkish populations, confirming previous associations both within the human leukocyte antigen (HLA) region, by performing an extensive study of this locus, and outside this region, such as *IRF5-TNPO3* and *NFKB1*. We also identified a suggestive association within the *GOT1-NKX2.3* locus, suggesting *NKX2.*3 as a potential candidate gene in SSc. In addition, we also studied the shared genetic component between SSc and other immune-mediated disorders through cross-disease meta-GWAS approach. Using this strategy, we found four new loci shared with Crohn's disease (*STAT3*, *IRF1*, *IL12RB2,* and *ZBTB9-BAK1*), and identified the IL-12/IL-23 signaling as one of a most common relevant pathway for both diseases. Furthermore, we analyzed the genetic component shared among four systemic seropositive rheumatic diseases (SSc, rheumatoid arthritis, systemic lupus erythematosus and idiopathic inflammatory myopathy) identifying 26 genome-wide significant common loci for at least two conditions, of which *NAB1*, *DGKQ*, *KPNA4-ARL14*, *LIMK1,* and *PRR12* had not been reported before, as well as determining that the type I IFN signalling pathway and its regulation play a more prominent role in these disorders.

However, one of the main limitations of GWAS is the difficulty to identify true causal genes, variants and cell types. Thus, we performed functional genomics studies including expression quantitative trait locus (eQTL) analysis and chromosome conformation capture studies (promoter capture Hi-C, pCHi-C) in order to provide a mechanistic link between non-coding SSc-associated variants and their target genes. Through the integration of GWAS and RNA-seq data we performed the first eQTL analysis in SSc, revealing that more than half of the eGenes detected were associated with the most important SSc hallmarks and highlighting the crucial role of the apoptotic process in SSc. On the other hand, pCHi-C analysis performed in $CD4^+$ T cells and $CD14^+$ monocytes from SSc patients and healthy controls, revealed cell-type specific interactions between SSc-associated loci and previously confirmed causal genes, such as *IRF8* in $CD14^+$ monocytes, and *CD247* and *STAT4* in $CD4^+$ T cells, as well as new potential candidate genes, especially *CXCR5,* which plays an important role in the differentiation of follicular helper T cells and has been associated with other ADs.

Finally, drug repurposing analyses performed throughout the different conducted studies identified more than 20 drug target genes already targeted in similar immune-mediated diseases, thus contributing to the potential repositioning of different drugs for its use in systemic sclerosis treatment.

RESUMEN

La esclerosis sistémica (SSc) es una enfermedad autoimmune reumática compleja, que presenta un fuerte componente genético y ambiental. Esta enfermedad está caracterizada por la presencia de fibrosis que puede afectar a la piel y a órganos internos, un fuerte daño vascular y una respuesta inmunológica alterada (incluyendo la presencia de auto-anticuerpos). Además, la SSc muestra un amplio rango de manifestaciones fenotípicas, así como características clínicas muy heterogéneas, lo que dificulta su correcto tratamiento.

La presente tesis doctoral se centró en el estudio del componente genético subyacente a la SSc. Hasta el momento, se han identificado más de 25 loci asociados con la susceptibilidad a desarrollar SSc mediante los llamados estudios de asociación de genoma completo (GWAS). Sin embargo, la mayoría de estos estudios han sido realizados en población caucásica. Con el objetivo de conocer mejor las bases genéticas de la SSc en otras poblaciones, realizamos un GWAS en pacientes con SSc de origen iraní y turco, confirmando las asociaciones previamente descritas, tanto dentro de la región del antígeno leucocitario humano (HLA), mediante un estudio en profundidad de la misma, como fuera de dicha región, incluyendo *IRF5-TNPO3* y *NFKB1*. Además, identificamos una asociación a nivel sugestivo del locus *GOT1-NKX2.3*, apuntando a *NKX2.*3 como un gen candidato potencial en SSc. También quisimos estudiar las bases genéticas compartidas entre la SSc y otras enfermedades inmunomediadas a través de un meta-análisis de datos de GWAS. Mediante esta estrategia, encontramos cuatro nuevos loci de susceptibilidad compartidos con la enfermedad de Crohn (*STAT3*, *IRF1*, *IL12RB2* y *ZBTB9-BAK1*) e identificamos la ruta de señalizacion IL-12/IL-23 como una de las principales vías patogénicas comunes a ambas enfermedades.

Además, se analizó el componente genético compartido entre cuatro enfermedades reumáticas seropositivas (SSc, artritis reumatoide, lupus eritematoso sistémico y miopatía inflamatoria idiopática), identificándose 26 loci de riesgo comunes a al menos dos enfermedades, de los cuales *NAB1*, *DGKQ*, *KPNA4-ARL14*, *LIMK1* y *PRR12* no se habian descrito previamente y destacando la señalización del interferón tipo I como una de las vías comunes de mayor relevancia entre estas cuatro enfermedades.

Sin embargo, una de las principales limitaciones de los GWAS es la dificultad para identificar los genes, variantes, o tipos celulares causales del fenotipo con el que se asocian. Por ello, decidimos realizar estudios de genómica funcional, incluyendo el análisis de locus de carácter cuantitativo de expresión (eQTLs) y estudios de captura de conformación de la cromatina (promoter capture Hi-C, pCHi-C) con el fin de relacionar las variantes asociadas con susceptibilidad a desarrollar SSc con sus genes diana. A través de la integración de datos GWAS y de secuenciación de ARN, realizamos el primer análisis de eQTLs descrito en SSc, observando que más de la mitad de los eGenes detectados están asociados con los principales rasgos característicos de la SSc, destacando el papel fundamental del proceso de apoptosis. Por otro lado, los análisis de conformación de cromatina pCHi-C realizados en linfocitos T CD4^{+} y monocitos CD14^{+} de pacientes con SSc y controles sanos, revelaron la existencia de interacciones específicas de tipo celular entre loci asociados a la SSc y genes causales previamente confirmados, como es el caso de *IRF8* en monocitos CD14^{+}, y *CD247* y *STAT4* en linfocitos T CD4^{+}, así como genes candidatos potenciales. Entre estos genes, destaca *CXCR5,* que juega un papel importante en la diferenciación de linfocitos T cooperadores foliculares y ha sido previamente vinculado con otras enfermedades autoinmunes.

Por último, realizamos un análisis de reposicionamiento de fármacos en varios estudios incluidos en la presente tesis, gracias a los cuales se identificaron más de 20 genes con diana farmacológica que, actualmente, están siendo abordados en otras enfermedades inmunomediadas similares, lo que indica que estos fármacos podrían ser potencialmente útiles para el tratamiento de la SSc.

INTRODUCTION

1. Systemic sclerosis: the complexity of autoimmunity

The immune response is a natural process by which our body defends itself against any pathogenic agent that can alter its homeostasis. However, this process can be altered, leading to an exacerbated response in the presence of a non-pathogenic autoantigen, triggering the autoimmunity process. This process is characterized by the loss of tolerance of T and B lymphocytes, the line of defense of adaptive immunity, against autoantigens of healthy tissues, which it recognizes as pathogens, leading to their deterioration (1).

Aberrant responses against self are implicated in more than 80 inflammatory disorders, the so-called autoimmune diseases (ADs), encompassing a huge spectrum of mostly unknown etiology. The autoreactivity of these disorders can range from the presence of circulating autoantibodies and minor tissue infiltrates to a strong pathogenic autoimmunity with immune-mediated organ injury. Unfortunately, the majority of ADs are debilitating, chronic and have no cure. Overall, these diseases present a high prevalence (7 - 9 %) in the population with a non-uniform distribution, preferentially affecting women, and varying between ethnic groups (1–3). This, together with their significant morbidity and mortality, the high medical cost to society, and their effect on the quality of life of patients, make these disorders one of the main challenges for research.

Most ADs are complex and clinically heterogeneous pathologies, strongly influenced by environmental and genetic factors, with a considerable epidemiological variability ranging from common (such as type 1 diabetes or rheumatoid arthritis) to rare diseases (systemic sclerosis or Sjogren's syndrome) (4). Based on the extent of tissue affected, these diseases can be classified into organ specific (such as type 1 diabetes or multiple sclerosis) and systemic (for example, rheumatoid arthritis or systemic sclerosis). Even when most of ADs are defined as complex diseases with polygenic nature, some of them are classified as monogenic diseases, with mutations in genes like *AIRE*, *FOXP3*, *IFIH1*, *DNASE1*, *TREX1*, *C1Q*, or *C4A,* to name but a few, which helped to the comprehension of the genetic bases underlying ADs.

Systemic sclerosis or scleroderma (SSc) is defined as a complex chronic AD that affects the connective tissue, characterized by an immune imbalance, vascular alterations, and an excessive collagen deposition leading to fibrosis of the skin and internal organs (5,6) (**Figure 1**). Immune imbalance in SSc consists of lymphocyte activation, leading to autoantibody production, excessive levels of pro-inflammatory cytokines and chemokines, and the dysregulation of the innate immune response. The most common symptom regarding vasculopathy in SSc patients is called Raynaud's phenomenon, a fibrointimal proliferation of small vessels and vasospastic episodes triggered by factors such as cold or stress, that can lead to tissue ischemia. Raynaud's phenomenon is one of the earliest clinical sign of SSc, however, this

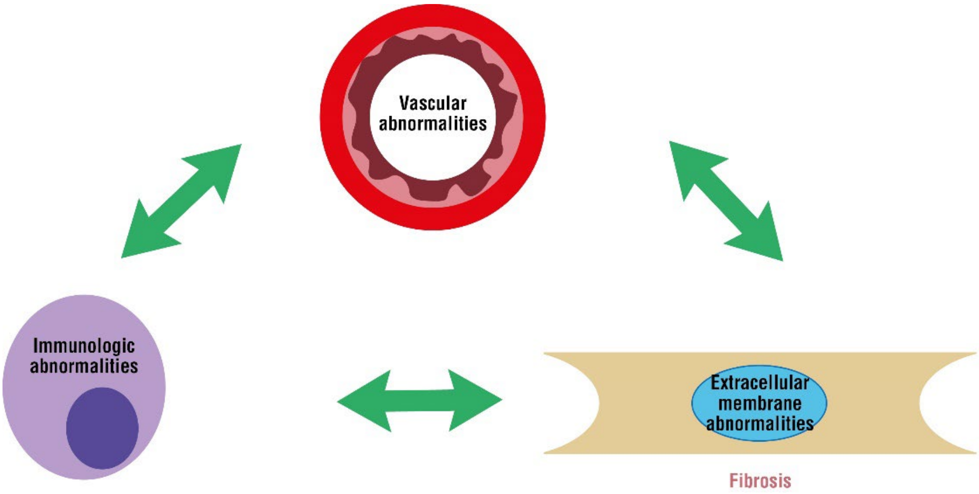

Vascular abnormalities
Immunologic abnormalities
Extracellular membrane abnormalities
Fibrosis

Figure 1. Overview of Systemic sclerosis pathogenesis.

microvasculature affection can evolve to such as SSc-related renal crisis (SRC) and pulmonary arterial hypertension (PAH), being this lung affection, along with pulmonary fibrosis (PF), the leading cause of death (7).

Endothelial cells (EC) are the main cell type implicated in these pathological processes, whose damage results in the activation of inflammatory cell infiltration and fibrotic processes, leading to vascular remodeling and irreversible structural changes (8–10) (**Figure 2**). In this line, EC apoptotic markers are elevated in the sera of SSc patients, leading to the recruitment of inflammatory cells such as fibroblasts and myofibroblasts. The etiology of this initial vascular injury is still unknown, but different factors, such as autoantibodies, viral infections, and the presence of toxins and oxidative stress, are in the spotlight (6). Interestingly, infiltration of inflammatory cells occurs more frequently at early stages of SSc and it reduces as the fibrotic process emerges. The fibrotic process of SSc is characterized by accumulation of fibrous extracellular matrix (ECM) composed of collagen, elastin, fibronectin and glucosamine, ultimately leading to loss of organ function. Another characteristic of this specific fibrosis is the presence of α-smooth muscle actin-positive (α-SMA), apoptosis-resistant myofibroblasts in the infiltrated tissue (11,12). Initially, it was considered that myofibroblasts were derived exclusively from an expansion of resident tissue fibroblasts; however, over the years, it has been observed that these cells are derived from many sources, including the transformation of adipocytes, activation of perivascular pericytes, and trans-differentiation of epithelial and vascular endothelial cells (13,14). In this sense, transforming growth factor beta (TGF-β) is considered the most significant growth factor that is able to trigger these epithelial- and endothelial-to-mesenchymal transitions, acting as a key regulator of fibrosis (15,16). Different studies have reported deregulated levels of TGF-β in skin and lung tissue from SSc patients (17,18), as well as its role in SSc pathogenesis in different *in vitro* and *in vivo* studies (19). As an example, a TGF-β receptor kinase inhibitor blocked bleomycin-induced lung disease in mice (20) and, in another study, the administration of a biological inhibitor of TGF-β was found to block SSc-like graft versus host disease (21).

In addition, vascular repair barely occurs in SSc patients, as angiogenesis and vasculogenesis processes are defective (8,9). In this regard, despite a general increase in many angiogenic factors, the emergence of avascular areas becomes more common in later stages of the disease (22). Some of the pro-angiogenic factors that have been observed upregulated in SSc patients are vascular endothelial growth factor (VEGF), fibroblast growth factor-2 (FGF-2) or platelet-derived growth factor (PDGF) (9). In fact, VEGF, which plays a central role in development of the blood and lymphatic vascular system, is totally disrupted in SSc patients (9,23).

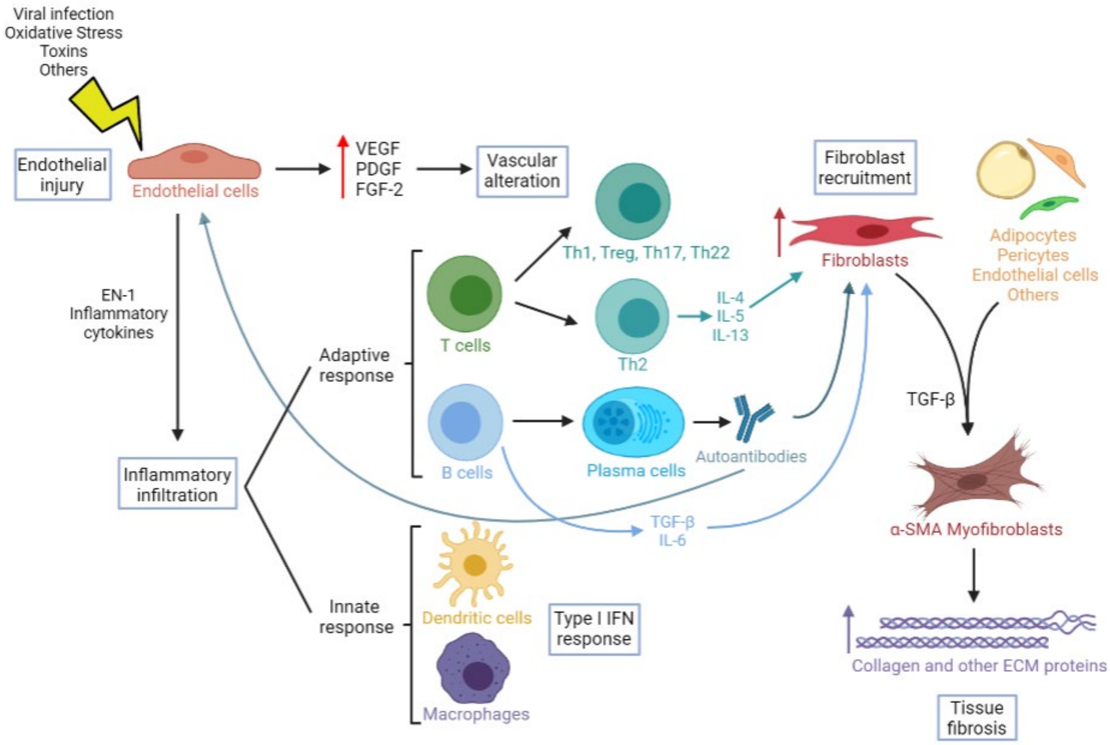


Figure 2. Summary of pathogenic mechanisms involved in systemic sclerosis.

Cross-talk between immune cells and stromal fibroblasts has long been considered a major driver of SSc pathogenesis and progression, occurring through the release of cytokines or facilitated by direct cell-cell contact (24). Both the innate and adaptive immune system are accepted to play a fundamental role in SSc pathogenesis (**Figure 2**). In this regard, it is not surprising that immune cells in SSc patients show specific characteristics. Lymphocytic infiltration of affected tissues has been observed in the earlier stages of disease, in which T cells show an activated phenotype, being also found in increased numbers in peripheral blood (25,26). In addition, there are many other infiltrated cell types, including dendritic cells and macrophages that, interestingly, show an upregulation in type I interferon (IFN) signaling, influencing adaptive immune response. Studies measuring the IFN signature from whole blood and peripheral blood show a large percentage of SSc patients with type I IFN excess (27–29). T cells are the main cell type from immune adaptive response involved in SSc pathogenesis. A variety of studies have implicated different CD4+ T cells subsets, such as T helper 1 (Th1), Th2, Th17, Th22, regulatory T (Treg) cells and T follicular helper cells, as well as CD8+ T cells (24). In this regard, Th2 cells are the most common subset identified in SSc infiltrate, characterized by the production of profibrotic mediators, such as interleukin (IL)-4, IL-5 and IL-13, which may interact with SSc fibroblasts inducing fibrosis (25). In fact, a strong imbalance between Th1/Th2 response is classically identified in SSc, observing increased levels of Th2 respecting Th1 cytokines (TNF-α, IFN-γ, IL-1 or IL-2). Although the paradigm of Th1/Th2 polarization has long been appreciated, an imbalance of Treg/Th17 has become increasingly important in recent years. In this sense, some studies indicate that Treg cells can be transformed into proinflammatory Th17 in the presence of TGF-β, IL-2 or IL-1β, which could explain this imbalance (30). Thus, the re-establishment of Th1/Th2 and Treg/Th17 balances may lead to the appearance of double-positive $CD4^{+}CD8^{+}$

T cells that have also been reported to play an important role in SSc pathogenesis, showing high levels in skin lesions of SSc patients and secreting very large quantities of IL-4 (31). On the other hand, the role of B cells in the pathogenesis of SSc has become increasingly apparent in recent years (32). Here of, B cells in SSc are hyperactive, observing increased levels of B cell activating factor (BAFF) in patients (33). These activated B cells can produce profibrotic cytokines, such as TGF-β and IL-6, and induce dendritic cell maturation, thus promoting profibrotic Th2 differentiation (32).

SSc can affect multiple organs. Along with lungs, gastroesophageal tract is the most affected tissue in SSc patients, but also the kidneys and heart are usually affected. Any area of the gastrointestinal tract can be affected, although esophageal involvement is the most common (5). On the other hand, lung involvement (both PAH and PF or interstitial lung disease, ILD) constitutes the leading cause of death in SSc, followed by SRC (5). The frequency of cardiac involvement is probably underestimated, certainly contributing to sudden death in patients associated with undercurrent sepsis (34). Other non-lethal manifestations of SSc include telangiectasia, calcinosis, Raynaud's phenomenon, digital ulcers, fatigue, and musculoskeletal and other chronic pain syndromes.

Although SSc is described as a heterogeneous disease, patients are usually stratified into two major clinical forms based on the extent of skin involvement: limited cutaneous (lcSSc) and diffuse cutaneous (dcSSc) disease. lcSSc fibrosis affection is restricted to hands, forearms, face and feet, and is characterized by a slower disease progression, with Raynaud's phenomenon appearing after several years. This subtype is also characterized by high ratio of pulmonary hypertension affection as well as the appearance of the CREST (calcinosis, Raynaud's phenomenon, esophageal dysmotility, sclerodactyly, and telangiectasia) syndrome, being also the most prevalent form affecting approximately 65% of patients. On the other hand, dcSSc have an extensive affection and a more aggressive and generalized fibrosis course, with an early onset of Raynaud's phenomenon and capillary destruction. In addition, this form of the disease is not limited to skin, affecting other visceral organs. dcSSc is associated with higher mortality rates, with a percentage of survival of 15% in 12 years as compared to 50% in the case of lcSSc (35–37).

The presence of autoantibodies is also a major SSc hallmark, and a numerous list of them have been described in SSc over the past decades (38). Nevertheless, only antinuclear antibodies (ANAs) are included in the classification criteria for SSc, as >90% of SSc patients present at least one of the three major subtypes, namely: anti-topoisomerase 1 antibodies (ATAs), anticentromere antibodies (ACAs), and anti-RNA polymerase III antibodies (ARAs). The remaining patients (3-11%) present rare autoantibodies or are negative for the presence of these three main subtypes. ANA patterns are generally mutually exclusive, allowing patients to be stratified early and providing the basis to manage a stratified approach (39). The prevalence of ATAs or anti-Scl-70 antibodies in SSc patients is reported to be between 20-30%. Furthermore, increased levels of ATAs are mainly associated with dcSSc and severe organ involvement, such as ILD (40,41). On the other hand, ACAs prevalence rounds about 20-40% and are more specific for lcSSc, being also associated with a longer disease duration and a better prognosis, as well as a higher risk of PAH (42,43). ARA+ prevalence is quite lower, rounding 10-20%. ARAs are developed in around 20% of dcSSc patients, and its presence indicates a high risk of rapid progression of skin thickening and SRC development (44–46).

SSc presents a high mortality rate, greater than any other rheumatic disease, with a high unmet medical need (5). It is considered a rare disease with a prevalence that varies substantially around the world, ranging from 7 to 700 cases per million, showing interstudy discrepancies (47). In this regard, lower estimated prevalence (<150 per million) have been observed in Northern Europe and Japan, while higher estimates are observed in Africa, Southern Europe, North America and Australia (276-443 per million) (48). Additionally, a North-to-South gradient has been reported in the European population (49,50). The risk to suffer SSc is also ethnic dependent, affecting more to black populations and Asians as compared to whites (47,51). Interestingly, a Native American tribe, the Choctaws, present the highest SSc prevalence described to date (660 per million) (52). It is also worth mentioning that, as occurs in other ADs and connective tissue disorders, the development of SSc is sex dependant and is much more common in women than in men, ranging from 3:1 to 12:1 (53). Nevertheless, male patients show a higher age-adjusted mortality in most studies (47).

2. Environmental component of systemic sclerosis

As previously stated, the etiology of systemic sclerosis is complex and still remains unclear. In this sense, it is thought to be caused by environmental factors influencing genetically susceptible individuals, thus triggering the onset of the disease and affecting its progression and severity (**Figure 3**). In this section we will review the current knowledge of environmental factors, as genetic component will be deeply addressed in further sections.

Little is known about the environmental, life-style and dietary factors that could trigger the disease, and its importance in the onset of SSc is still not robustly established due to methodological limitations, such as small sample size. Nevertheless, the main environmental factors known can be divided into three categories: (A) chemical agents, including particularly occupational agents; (B) biological agents, such as infections and dietary contaminants; and (C) physical agents, including ultraviolet and ionizing radiation, as well as electric and magnetic fields.

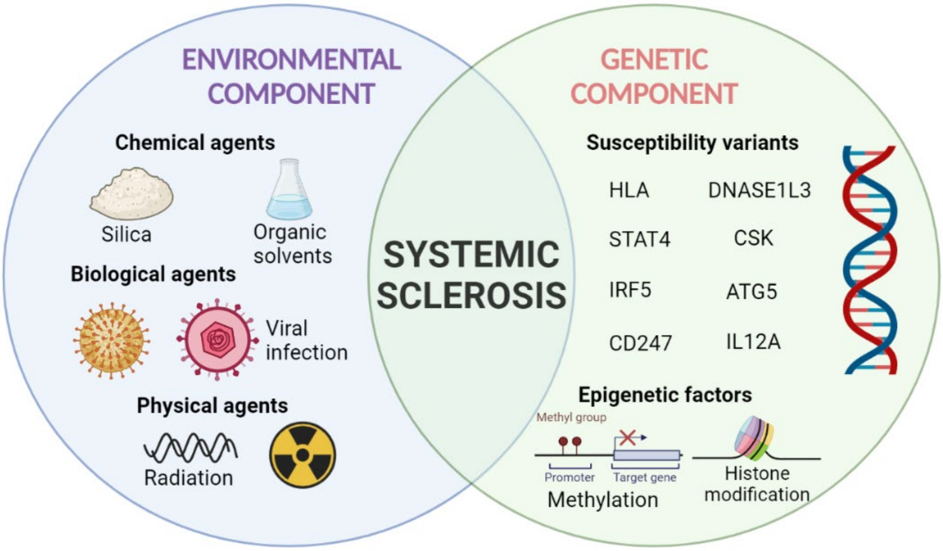


Figure 3. Summary of etiological factors influencing systemic sclerosis, determined by environmental and genetic components.

2.1. Chemical agents

Silica dust released from fractured silica crystals was one of the first agents associated with the development of SSc. First reports identified SSc clusters in Scottish stonemasons date back to 1914 (54). Further studies performed years later observed similar patterns in South African gold miners, and North American coal miners, pointing to crystalline silica as the causal factor (55,56), In this regard, a more recent study performed by Haustein *et al* (57) points to 25 to 50 times increased SSc risk in individuals exposed to crystalline silica than those not exposed. In spite of its long known association with SSc, pathological mechanisms underlying this association remain unclear. Nevertheless, it is known that silica acts as a strong T cell adjuvant, and thus, could lead to tissue damage and inflammatory response in genetically predisposed individuals (58). Additionally, different studies indicated that silica administration leads to activation of T and B cells, autoimmunity related apoptosis, and fibroblast proliferation (59,60).

In addition, organic solvents have been reported to increase SSc risk. To date, their applications are increasingly diversified, with more and more occupations associated with the exposition to organic solvents (61). First associations date from the 1950s (62) and further meta-analyses performed in recent years, as the one reported by Kettaneh *et al* (63), showed an increased risk (almost double) to develop SSc in individuals exposed to organic solvents than in controls, being this risk greater in men than in women. Additionally, occupational exposure to solvents acts as a predictive parameter of SSc severity. In this regard, it has been observed that SSc patients exposed to organic solvents exhibited dcSSc and microangiopathy more frequently than non-exposed individuals (64). As with silica, the role of solvents in pathological pathways of SSc remains unclear. Nevertheless, it is thought that the linkage of organic solvents with nucleic acids and proteins could result in immune disruption, initiating cellular and humoral autoimmune responses and stimulating fibrogenic responses, leading to an increased risk of SSc (65).

The exposition of subjects to asbestos, used due to heat resistance occurring in construction and mining, is highly correlated with the risk of developing different ADs, including SSc (66). Particularly, individuals exposed to amphibole, a kind of asbestos, developed ANAs more often than non-exposed individuals (67). In this regard, intratracheal injection of amphibole asbestos in rodents showed an induced synthesis of ANAs and modifications in serum cytokines (68,69), observing increased concentrations of IL-17 triggered by other cytokines, such as IL-6 or TGF-β (70).

There are other industrial agents, such as welding fumes, epoxy resins, vinyl chloride, or formaldehyde that have been related to SSc. On the other hand, many other non-occupational chemical factors reported no association with SSc, such as drug consumption, implants (silicone, prosthesis, contact lenses), smoking, or dyeing hair (61).

2.2. Biological agents

Infections are the main associated factor within this group. Several infectious mechanisms have been described, triggering autoimmune response via EC damage, self-reactive antibodies and molecular mimicry processes (71). In this regard, parvovirus B19 has been speculated to play a role in SSc, as it has been detected in bone marrow biopsies of more than half of SSc patients, showing an increased prevalence in skin tissue from patients as compared to controls. This fact highlights the possibility that parvovirus B19 may play a role in the formation of skin tissue abnormalities in the disease (72,73). The herpesvirus family, including cytomegalovirus (CMV) and Epstein-Barr virus (EBV), are related to SSc onset due to their ability to infect fibroblasts and ECs, with consequent autoimmunity alterations via molecular mimicry (74,75) In this regard, IgG from SSc patients recognises the protein UL94 from CMV, inducing the apoptosis of ECs through its union with the EC surface integrin-NAG-2 protein complex (74). Remarkably, *in vitro* studies have reported the capacity of EBV to infect SSc fibroblasts persistently and to induce the dysregulation of the innate immune response (76). Moreover, *Helicobacter pylori* may be involved in the development of SSc through endothelial damage and vascular changes (71). Indeed, an increased prevalence of *H. pylori* has been observed in SSc patients as compared to healthy subjects (77,78).

Dietary factors can play a role in ADs development. One classical example could be celiac disease, produced by gluten ingestion. Nevertheless, based on currently published studies, there is no strong evidence to show that food and dietary contaminants play a role in SSc. Some of these studies include: exposure to complex food in infants, breast-feeding, alcohol consumption, or consumption of food chemicals, dyes and additives (61).

2.3. Physical agents

Physical agents, including ultraviolet and ionizing radiation, and electric and magnetic fields, are becoming a growing target as etiological factors of different ADs. In this sense, different authors observed an association between ultraviolet exposure and the risk of developing multiple sclerosis, as well as between ionizing radiation and the development of Grave's disease and autoimmune thyroiditis (57). However, no studies have been performed yet trying to assess the association of these factors with SSc.

3. The multifactorial genetic component of systemic sclerosis

The main risk factor associated with the development of SSc is the occurrence of the disease in a close relative. This fact highlights the deep importance of the genetic component in SSc. In this regard, having a sibling with SSc represents the major indicator of risk to suffer from the disease (with a 15-19 fold increase over the general population), and having other first-degree familiar suffering from SSc suppose a fold increase of 13-15 (79,80). Indeed, it has been detected a high autoantibody concordance, particularly in ANAs, between twins and in SSc multi-case families. On the other hand, low concordance was observed for SSc in these twin studies (81,82). In addition, as previously mentioned, ancestry has a relevant role in SSc susceptibility, which supports the role of the genetic component of the disease.

However, as it happens in other autoimmune and rheumatic diseases, the genetic component of SSc is not clear, as it is not inherited in a mendelian fashion. As a complex disease, the estimation of SSc heritability is difficult, and its real importance in the development of the disease remains controversial. In order to estimate missing heritability in complex diseases, Lee *et al* (83) developed the GREML method, which is based on the assumption that more genotype sharing between non-related subjects should result in a greater phenotypic concordance. Through this methodology, our group performed a study in which the SSc estimated heritability on the observed scale (h_o^2), defined as the proportion of variance explained in case-control studies by associated genetic variants, was of 0.39 and 0.44 in two different cohorts (84). It is worth mentioning that significant genetic variants associated with SSc only account for ~20% of the estimated SSc heritability, which indicates that there are still other loci to be associated with SSc.

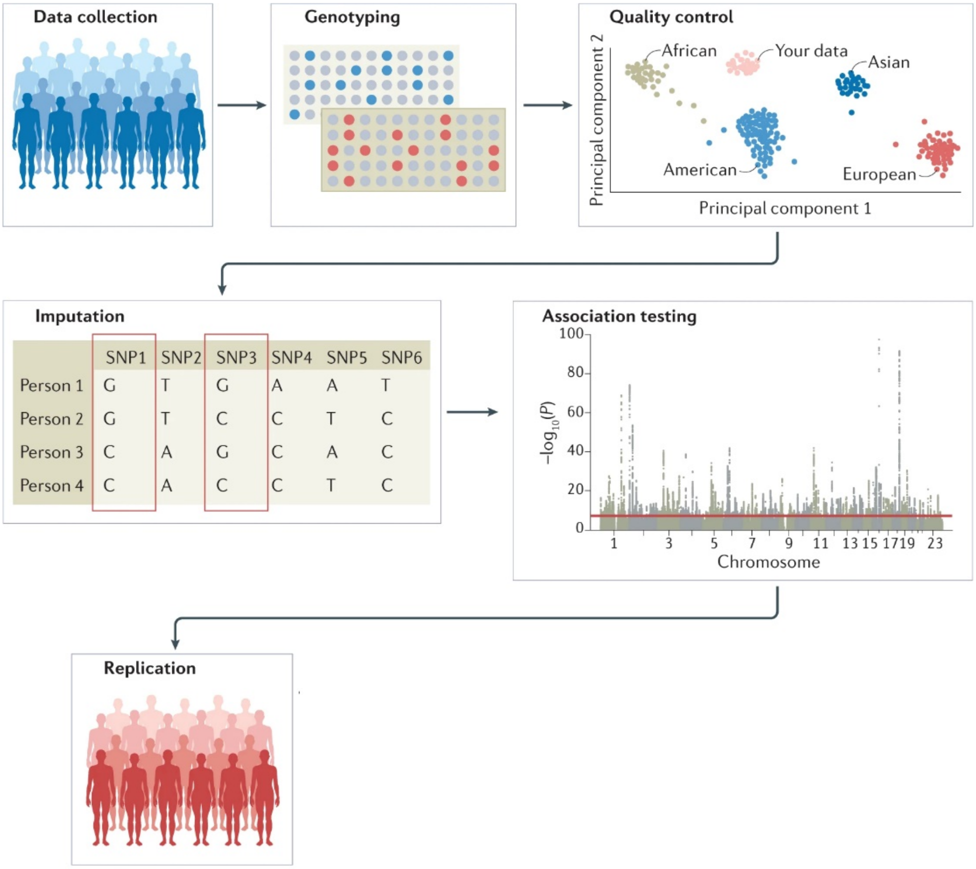
On the other hand, over the past few years, other heritable changes that influence gene expression without altering the DNA sequence have been discovered. These epigenetic factors corresponding to the interaction of environmental and genetic components are known to contribute to the risk of SSc. The knowledgement acquired about the complex genetic component of SSc, from the first candidate gene studies to the last epigenetic analyses will be afforded in this section.

3.1. Genotyping studies

As a complex disease, first studies performed in order to unravel the genetic component of SSc were based on the pursuit of genetic markers that may be associated with the disease. In this sense, single nucleotide polymorphisms (SNPs) are the most studied genetic markers. SNPs are changes in single base pairs of the DNA sequence occurring between individuals that commonly have two alleles. The minor allele frequency (MAF) must be above 1% in the overall population to be defined as SNP. These mutations occur once every 1000 nucleotides on average, meaning 4 to 5 million SNPs in a single individual. To date, more than 80 million SNPs have been discovered in different populations (85). Variants with a MAF below 1% would be considered as rare variants.

In order to study the association of these genetic marks with a certain phenotype, it is necessary to perform case-control studies. These are epidemiological studies in which cases (individuals affected by a specific phenotype) are compared to controls (which are non-affected individuals). In this specific case, the frequency of the minor allele of a SNP is compared between cases and controls and, if the difference is statistically significant, the SNP is considered to be associated with the disease. This kind of studies can be designed based on previous knowledge about the disease, by analyzing selected polymorphisms from interesting regions, the so-called candidate gene studies, or can be hypothesis free, in which a huge quantity of SNPs are interrogated. Candidate-gene studies were first developed, starting with the selection of a particular locus or polymorphisms based on the possible functional implication of a position or region in SSc. In this regard, variants located in coding or regulatory regions are preferentially studied over non-coding regions. Chose targets are also based on their association with other ADs, autoimmunity, inflammation, fibrosis, or vascular function in previous studies. Thus, the selection of candidate genes is determined by previous publications and partial scientific knowledge, and are directed by the researchers. However, candidate gene studies can be a powerful tool to analyze the contribution of determined loci, especially when the cohorts analyzed are large enough.

In the specific case of SSc, initial candidate gene studies comprised small cohorts, leading to low statistical power, and thus few genetic susceptibility loci were identified apart from the human leukocyte antigen (HLA) region, the strongest and best-known region associated with ADs. These associated regions correspond with *IRF5* and *STAT4*, being some of the firmest susceptibility regions identified outside the HLA (86). On the other hand, the advances in genotyping technologies and the recruitment of ever increasing cohorts of patients and controls, as well as the cheapening of these technologies, led to the possibility of genotyping a considerable amount of SNPs simultaneously (87). In this regard, genome-wide association studies (GWAS) performed in well-powered cohorts became a revolution to find plenty of associations across the genome with different diseases. The design of the genotyping arrays used to perform these analyses are based on the known linkage disequilibrium (LD) patterns of the genome and haplotype structure, in order to optimize the number of variants to be included with the maximum coverage by genotyping a few hundred thousand of variants. This strategy presents a series of advantages as compared to previous candidate gene association studies: it is a hypothesis-free approach, in which the associations observed are not focused on a specific region of interest, and at the same time, it is a hypothesis generating approach, since the novel discovered loci may involve new pathways that can be studied in further analysis. Another methodology that has improved the technology of GWAS is the genotype imputation process. Through these imputation algorithms, scientists are capable of inferring non-genotyped variants (missing genotypes) based on nearby observed genotypes, by comparing them with haplotypes of reference panels (88). The number of SNPs tested for association is now increased through these *in silico* genotypes, which improves the power of the study and facilitates meta-analyses and the ability to fine-map or identify causal variants. Nevertheless, due to the large number of tests performed, a restrictive multiple testing correction threshold must be taken into account in order to avoid false-positive associations. In this sense, the standardized significance threshold for GWASs is established at *p*-value $< 5x10^{-8}$, corresponding to Bonferroni correction for one million independent tests (SNPs) (89). The replication of the results obtained in GWASs in independent cohorts is mandatory, as the associated variants tend to show inflated effects (**Figure 4**).



	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
Person 1	G	T	G	A	A	T
Person 2	G	T	C	C	T	C
Person 3	C	A	G	C	A	C
Person 4	C	A	C	C	T	C

Figure 4. Overview steps for conducting GWAS (Modified from Uffelman *et al*, Nat. Rev. Methods Primers. 2021)

The first SSc GWAS was performed in 2009 in Korean population. Nevertheless, due to a low sample size, only SNPs in the *HLA-DPB* region were associated with SSc susceptibility, specifically with ATA+ and ACA+ patients (90). In 2010, a much larger GWAS performed in European populations, in which our group was involved, was published (91). In this study, *CD247* was identified as a new associated locus in SSc, also confirming other previously associated loci*,* including *IRF5*, *STAT4,* and the HLA region, at the genome-wide significance level. Notably, *CD247* was robustly replicated in French population in a subsequent study (92). In 2011, a third GWAS performed in SSc patients and controls from France identified *TNIP1*, *PSORS1C1* and *RHOB* as novel risk loci in SSc (93). However, in a later replication study, our group confirmed *TNIP1* as a risk locus for SSc, but failed to replicate the reported associations within *PSORS1C1* and *RHOB*, thus highlighting the importance of including replication cohorts and a large enough sample size in this kind of studies (94). On the other hand, in these published GWASs, many SNPs did not reach the genome-wide significance level but remained in the grey zone, defined as SNPs with a *p*-value between 5x10^{-5} and 5x10^{-8}. In this regard, subsequent follow-up studies have helped to identify new associations by focusing on regions of interest that could reach significance when they are approached independently. Through this focus on certain regions, it is possible to analyze specific signals in larger cohorts without an extra economic cost. This approach has been applied in SSc, successfully identifying new associated loci at the genome-wide level of significance, such as *IL12RB2*, *CSK*, *PPARG*, *IL12RB1* and *TYK2* (95–98).

Even when the design of GWAS arrays include an increasing number of SNPs, most of them have modest effect sizes. This missing heritability is characteristic from highly polygenic and complex diseases, such as SSc, which can be affected by different rare variants. In this sense, odds ratio (OR) for most of the SNPs range from 1.1 to 1.4 approximately, explaining just a small proportion of the disease (99). It is proposed that missing heritability could lie in specific regions of the genome that may not be well covered by GWAS arrays, or could belong to rare variants with a large effect size that are difficult to identify. With the purpose of overcoming this problem, the Immunochip array was created. This custom genotyping platform contains almost 200,000 variants, including SNPs and other insertion-deletions, of particular interest in immune-mediated diseases, with a dense coverage of 186 autoimmunity loci and the HLA region (100). In addition, the price of Immunochip is much lower than most of the GWAS chips, enabling groups to finance genotyping of very large cohorts. Thanks to this, Immunochip platform has significantly contributed to the discovery of many genetic loci in different ADs. Particularly in SSc, the first Immunochip study, published by our group, identified three new susceptibility loci for SSc, including *DNASE1L3*, *IL12A,* and *ATG5*, implicating new pathological pathways for the disease, such as autophagy or apoptosis (101). In this sense, thanks to the great coverage of the HLA region, our group was able to further dissect the long-known association between this region and SSc, highlighting six polymorphic amino acid positions in HLA-DRB1, HLA-DPB1, and HLA-DQA1, and seven SNPs independently associated (101). Another Immunochip study was published the same year including data from an Australian cohort, confirming part of the reported associations observed in our study (102).

It is worth noting that despite the advances obtained in the understanding of the genetic basis of SSc thanks to GWAS and Immunochip, the number of well-established loci associated with the disease was relatively low (15 loci) (103) as compared with other systemic ADs such as rheumatoid arthritis (RA) or systemic lupus erythematosus (SLE), reaching more than 100 and 80 associated loci, respectively (104,105). These differences are due to the low prevalence of SSc which makes it difficult the recruitment of large cohorts and, consequently, the achievement of enough statistical power to detect small association signals. To partially overcome this problem, our group recently published the largest GWAS in SSc performed to date, in which 14 independent European cohorts were meta-analyzed. Almost 10,000 SSc patients and 18,000 controls were included in the analysis, reaching a total of 27 independent associated signals, including 13 new risk loci (106). Some of the prioritised genes such as *DDX6* or *FLNB* highlight the importance of fibrotic and vasculopathy pathways in the pathogenesis of the disease. In addition, a posterior fine-mapping study of the HLA region performed in the same cohorts highlighted nine SNPs, nine classical alleles, and seven amino acids that modelled the observed associations with SSc, being the largest HLA analysis performed to date in SSc (107). This study confirmed the association of the two main HLA class II classical alleles related with SSc, *HLA-DRB1*11:04* and *HLA-DPB1*13:01*, and revealed a novel association within HLA class I corresponding to *HLA-B*08:01* (107).

In recent years, meta-analyses of GWAS (meta-GWAS) have facilitated the discovery of tons of new susceptibility loci, as they combine data from multiple studies of relatively small sample size, thus increasing the statistical power and the chance to identify significant associations (108). One of the most interesting uses of meta-GWAS in autoimmunity is the search for its shared genetic component. Scientific evidence has demonstrated that more than half of genome-wide significant AD-associations are shared by at least two distinct ADs (109,110). In this regard, cross-disease meta-GWAS, which combine GWAS data from different diseases as a single phenotype, have identified new susceptibility loci shared between SSc and SLE (111) and between SSc and RA (112). Furthermore, this approach has been applied to combine genomic data from multiple immune-mediated diseases in the same meta-analysis, leading to the discovery of more than 30 new shared susceptibility loci (113–115). During the period of this thesis we have applied this approach to identify the genetic overlap between SSc and Crohn's disease

(CD) in a study comprising more than 5,700 SSc patients, 4,500 CD patients, and 14,500 healthy controls (116). Even though SSc and CD present apparently unrelated phenotypic traits, several lines of evidence support the existence of a shared genetic component between them. First of all, results from large-scale genetic studies performed in each individual disease have shown a genetic overlap between SSc and CD, with several genetic risk loci common to both conditions, such as *IRF8*, *TYK2*, *STAT4*, and *GSDMA*/*IKZF3,* as well as the HLA region (106,117). Additionally, there is an important fibrotic component in both diseases. Even when fibrosis is one of the primaries hallmarks of SSc, it also appears in CD and is one of the main reasons that leads to a necessity of surgical intervention in the distal part of the small intestine (118,119). In this line, it has been observed an increased risk of idiopathic pulmonary fibrosis (IPF) in individuals affected by CD (120), being this fibrosis of the lungs one of the most common complications in SSc, leading to ILD (121). Furthermore, we applied a more complex approach including GWAS data from four systemic seropositive immune-mediated inflammatory diseases (IMIDs) including SSc, SLE, RA, and idiopathic inflammatory myopathies (IIM), comprising more than 11,600 cases and 19,700 healthy controls (122). These four systemic rheumatological IMIDs are heterogeneous diseases of the connective tissue that share clinical and epidemiological manifestations as well as life-threatening complications (123). The common genetic component of these conditions has not been previously assessed systematically, although the overlap of associated genes is elevated when performing a pairwise comparison (115). Autoantibody production is the main feature of these diseases, comprising additionally a broad deregulation of the innate and adaptive immune response. However, the low prevalence of most of these diseases hinders the collection of large datasets that makes possible to attain sufficient statistical power. Therefore, our study aimed to combine previously published GWAS datasets—all from European descent populations—to identify shared genetic aetiologies among systemic seropositive rheumatological IMIDs in a systematic fashion. Both cross-disease meta-GWAS have been included in the present PhD dissertation (116,122).

Despite the fact that GWASs have attempted to discover hundreds of susceptibility loci for different ADs, the large majority of research to date have been performed with samples of Caucasian or European-descent population (**Figure 5**). This European bias has important implications for risk prediction of diseases in different populations (124). In this regard, most of GWAS arrays as well as the Immunochip are designed for use in white European population, being less informative for other ethnic groups (100). In the case of SSc, some studies have described HLA region associations in different non-European populations such as African-americans, Koreans, or Mexicans, identifying

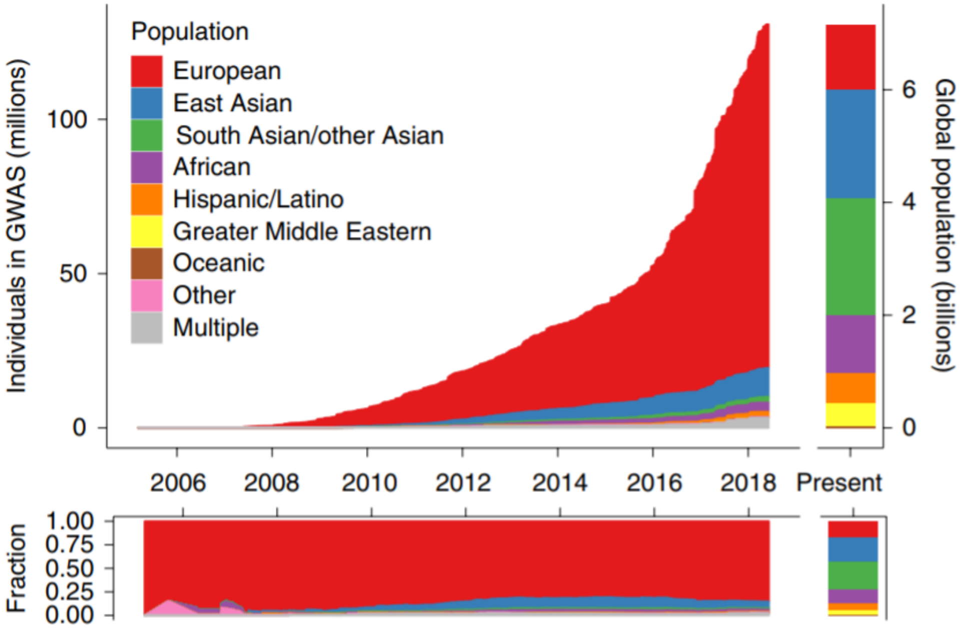


Figure 5. Ancestry of GWAS participants over time, as reported by GWAS catalog (199)(Extracted from Martin *et al*, Nat Genet. 2019)

different classical alleles associated with the disease depending on the population (90,125,126). In addition, the first transethnic meta-analysis published in SSc, including Japanese and European populations, identified two new non-HLA susceptibility genes: *GSDMA* and *PRDM1* (127). Thus, with the purpose to identify new susceptibility loci and discern the genetic landscape of SSc in different non-European populations, we decided to perform, for the first time, an extensive GWAS and HLA region analysis in SSc patients from Iranian and Turkish populations, including more than 800 cases and 1,400 healthy controls (128). The results of this study will be approached in the present PhD dissertation.

Thanks to the advances performed from early discoveries of familial aggregation to the latest large-scale cross-disease meta-analyses (**Figure 6**), a total of 30 loci outside the HLA region have been firmly associated to SSc to date (**Table 1**). Most of the SSc susceptibility loci robustly replicated are involved in innate and adaptive immune response, as well as in autophagy and apoptosis pathways. Regarding innate immunity, type I IFN signaling is the most overrepresented pathway, including four interferon regulatory factors (IRFs) (**Table 1**). On the other hand, several genes related with adaptive immune response are also overrepresented in SSc risk loci, including *TNFSF4*, which is involved in B and T cell proliferation and survival; or *CD247*, which

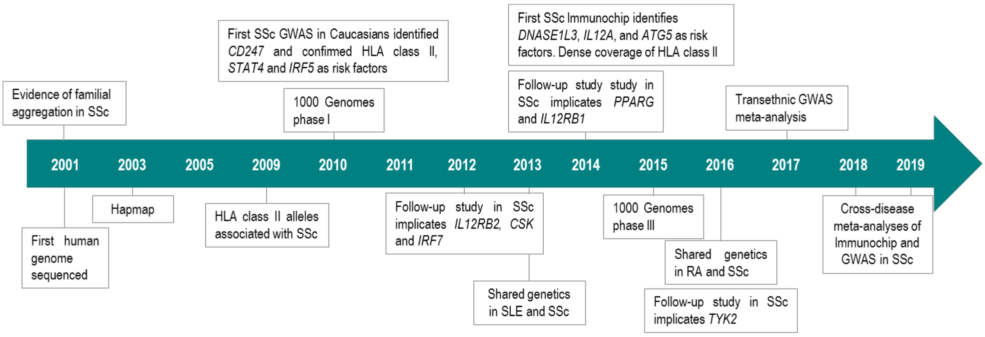

First SSc GWAS in Caucasians identified *CD247* and confirmed HLA class II, *STAT4* and *IRF5* as risk factors
First SSc Immunochip identifies *DNASE1L3*, *IL12A*, and *ATG5* as risk factors. Dense coverage of HLA class II
Evidence of familial aggregation in SSc
1000 Genomes phase I
Follow-up study study in SSc implicates *PPARG* and *IL12RB1*
Transethnic GWAS meta-analysis
2001
2003
2005
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
Hapmap
First human genome sequenced
HLA class II alleles associated with SSc
Follow-up study in SSc implicates *IL12RB2*, *CSK* and *IRF7*
Shared genetics in SLE and SSc
1000 Genomes phase III
Shared genetics in RA and SSc
Follow-up study in SSc implicates *TYK2*
Cross-disease meta-analyses of Immunochip and GWAS in SSc

Figure 6. Timeline of major advances made in genetics of systemic sclerosis (Extracted from Acosta-Herrera *et al*, Curr Rheum Rep. 2019)

forms part of the T cell receptor (TCR) complex. Furthermore, Jak/STAT and IL12/23 signaling pathways are also overrepresented in SSc risk loci, including genes such as *TYK2*, which encondes a tyrosine kinase that mediates signaling of IL-12 family cytokines; or *IL12A* and IL-12 receptor B genes (*IL12RB1*, *IL12RB2*) (**Table 1**). Other important processes implicated in SSc pathology such as autophagy and apoptosis, were also overrepresented at the genetic level. In this sense, *DNASE1L3* plays an important role in in DNA fragmentation during apoptosis, and *GSDMA/B* present an important role in pyroptosis, a form of cell death triggered by inflammatory signals.

As it has become clear, genotyping studies have been crucial to better understand the genetic background of SSc and other ADs. Nevertheless, most of the SNPs associated with the disease in large-scale genotyping studies map to non-coding regions of the genome, which are in fact enriched in regulatory marks, such as enhancer regions (129). Thus, the next step after defining a new locus is to identify the causal variants and the mechanism of action underlying those disease-associated variants, thus allowing the translation of genetic findings to the clinic (130). During the last five years, a paradigm shift has occurred, in which genotyping studies are being combined with different other technologies to create a multi-omic approach that better explains the genetic component of complex diseases, such as SSc, through a holistic view (131). In this regard, many different approaches have emerged along with genotyping studies that help elucidate the molecular basis of SSc, such as transcriptomics or different epigenomic studies (DNA methylation, histone modification, chromatin interaction, etc).

Table 1. Non-HLA susceptibility loci firmly associated with systemic sclerosis to date.

Locus	Chr	Gene name
Innate immune response		
IRF4	6	Interferon Regulatory Factor 4
IRF5-TNPO3	7	Interferon Regulatory Factor 5 Transportin 3
IRF7	11	Interferon Regulatory Factor 7
IRF8	16	Interferon Regulatory Factor 8
PRDM1	6	PR/SET Domain 1
TNFAIP3	6	Tumor necrosis factor Alpha Induced Protein 3
TNIP1	5	TNFAIP3 Interacting Protein
NFKB1	4	Nuclear Factor Kappa B Subunit 1
Adaptive immune response		
TNFSF4	1	Tumor necrosis factor Superfamily Member 4
CD247	1	T-Cell Receptor T3 Zeta Chain
CSK	15	C-Terminal Src Kinase
PTPN22	1	Protein Tyrosine Phosphatase Non-Receptor Type 22
STAT4	2	Signal Transducer and Activator Of Transcription 4
BLK	8	BLK Proto-Oncogene, Src Family Tyrosine Kinase
IL-12 Signaling Pathway and cytokines		
IL12A	3	Interleukin 12A
TYK2	19	Tyrosine Kinase 2
IL12RB1	19	Interleukin 12 Receptor Subunit Beta 1
IL12RB2	1	Interleukin 12 Receptor Subunit Beta 2
Apoptosis and Autophagy Pathways		
DNASE1L3	3	Deoxyribonuclease 1 Like 3
ATG5	6	Autophagy Related 5
RAB2A-CHD7	8	Member RAS Oncogene Family Chromodomain Helicase DNA Binding Protein 7
GSDMA	17	Gasdermin A
GSDMB	17	Gasdermin B
Vascular homeostasis, fibrosis and others		
PPARG	3	Peroxisome Proliferator Activated Receptor Gamma
NAB1	2	NGFI-A Binding Protein 1
DDX6	11	DEAD-Box Helicase 6
DGKQ	4	Diacylglycerol Kinase Theta
POGLUT1-TIMMDC1-CD80-ARHGAP31	3	Protein O-Glucosyltransferase 1 Translocase of Inner Mitochondrial Membrane Domain Containing 1 Cluster of Differentiation 80 Rho GTPase Activating Protein 31
TSPAN32, CD81-AS1	11	Tetraspanin 32 Cluster of Differentiation 81 Antisense RNA 1
NUP85-GRB2	17	Nucleoporin 85 Growth Factor Receptor Bound Protein 2

3.2. Transcriptomic studies

Genome-scale gene expression data allow us to infer, from messenger RNA (mRNA) expression measurements, different pathways that are implicated in a specific tissue, cell type, or other kind of samples. Microarrays were the first platforms used to scan the expression levels of thousands of genes at the same time from multiple affected tissues in SSc patients, including skin (132–134), lung (135,136), blood (137,138), or esophagus (139). One of the main features that can be drawn from these data is the enormous heterogeneity of molecular processes that are altered in SSc patients. Gene expression studies from skin biopsies revealed inflammatory and fibrotic signatures. Surprisingly, nearly identical gene expression patterns were observed in biopsies from lesional and non-lesional skin, highlighting the systemic nature of the disease in which non affected tissues can show aberrant gene expression (134). Similar results were observed in a comparison between upper and lower esophagus (139) A more recent study analyzed gene expression in multiple tissues through a novel multi-network approach, finding key similarities in the fibrotic- and immune-related expression patterns, and also implicating alternative macrophage activation in lung tissue (140).

Nevertheless, most of the SSc gene expression datasets are performed in whole tissue samples that are in fact composed of a mixture of different cell types, making it difficult to assess the specific role of a cell type in the pathogenesis of the disease. In this regard, different functional genomic networks have emerged in order to study the interactions of cell type- and tissue-specific genes (141). In this sense, the combination of gene expression data with other datasets is necessary to create functional genomic networks that can extract specific cell type signals and meaningful pathways from whole tissue studies. A study published by Mahoney *et al*, in which authors connected gene expression from three independent SSc skin datasets with susceptibility SNPs observed in GWAS, represents a good example of this (142). Interestingly, authors observed that risk SNPs identified in SSc were almost exclusively connected with gene expression data in the immune system context (142). This type of study represents a good approach to connect GWAS signals and transcriptomic data. Indeed, the analysis of these signals in the context of gene expression in cells or tissues has allowed a better understanding of human genetics through the study of expression quantitative trait locus (eQTLs) (143). Briefly, an eQTL is a locus that explains part of the genetic variance of a gene expression phenotype, involving a direct association test between genetic markers and gene expression levels. eQTLs are usually categorized depending on the distance between the associated SNP and the interacting gene, usually describing regulatory regions as *cis* (proximal), for those interactions within 1 Mb (megabase) distance, and *trans* (distant), for those at least 5 Mb distance or those occurring between different chromosomes (**Figure 7**). In this line, an eQTL analysis performed in monocyte-derived macrophages from SSc patients suggested that the contribution of the risk variant rs3894194, associated with SSc in a previous GWAS, can be mediated by *GSDMA* expression in macrophages, a gene implicated in pyroptosis, which is intimately related with inflammation processes (144).

On the other hand, the majority of gene expression analyses performed to date in SSc are based on microarray technology, which has a limited potential as compared with next-generation sequencing (NGS) techniques. Indeed, RNA sequencing (RNA-seq), based on NGS technology, presents a much higher resolution and deep-coverage, as well as high sensitivity at the extreme of gene expression ranges and low sensitivity to background noise, making it a much more accurate tool for transcriptomics studies (145). The largest RNA-seq profiling of SSc patients performed to date

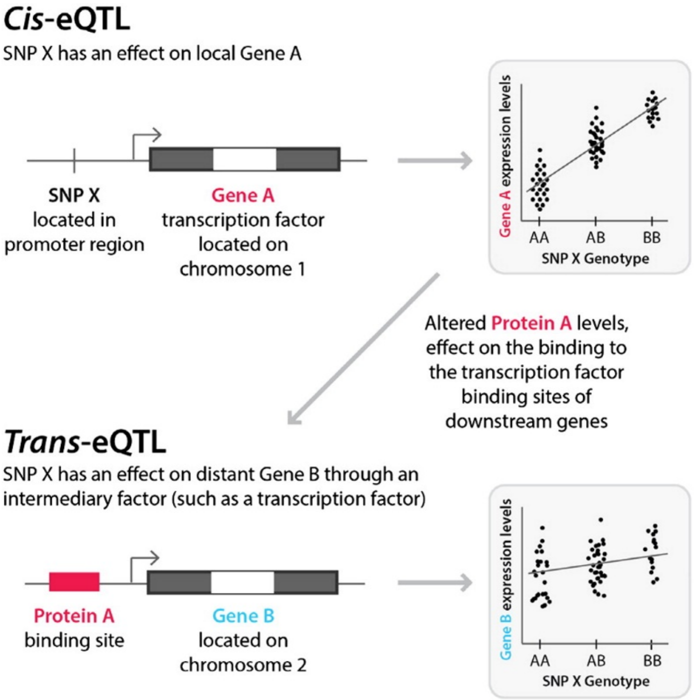


Figure 7. Example of eQTL action in *cis* and *trans*. (Extracted from Westra *et al*, Biochim Biophys Acta Mol Basis Dis. 2014)

was carried out in whole blood samples from a large well-characterized European cohort in the context of the PRECISE systemic autoimmune diseases (PRECISESADS) project, including a total of 162 patients and 252 controls (146). In this study, the analysis of differentially expressed genes between patients and controls indicated a deregulation of important pathways implicated in SSc pathogenesis, such as type I IFN, toll-like receptor cascade, and platelet degranulation and activation. Our group has taken part in the PRECISESADS project, in which different systemic ADs have been characterized at the molecular level, including genotyping and RNA-seq data. In this regard, we performed an eQTL analysis in more than 300 SSc patients and 500 controls, integrating GWAS and RNA-seq data in order to provide a mechanical link between SSc associated variants and their effect on gene expression (147). The results obtained in this study are part of the present

PhD dissertation.

It is also worth mentioning the contribution of the recently developed single-cell RNA sequencing (scRNA-seq) studies in revealing complex and rare cell populations in different ADs through the obtention of high-resolution sequencing on individual cells (148) In the specific case of SSc, a few studies have emerged exploring the main cell types associated with its pathogenesis: fibroblasts (149,150) and T cells (151). One of these studies revealed that *SFRP2/DDP4*-expressing progenitor fibroblasts are the main cell population that differentiates to pathogenic SSc myofibroblasts, driven by upstream transcription factors including *STAT1*, *RUNX1* or *IRF7* (149). Another study performed on CD4$^+$ T cells identified several subsets of tissue-resident and recirculating T cells in skin from SSc patients, highlighting a distinct CXCL13$^+$ T cell subset expressing a T follicular helper gene signature that promotes B cell responses within inflamed skin (151).

3.3. Epigenomic studies

The definition of what epigenetics is has broadly changed over the decades. Nowadays, it is usually defined as the study of heritable changes that affect gene expression independent of altering the nucleotide sequence itself. These crucial regulatory events are involved in how DNA is packed and how chromatin is structured, thus coordinating gene transcription during different physiological processes, but also in pathological processes. Common epigenetic mechanisms include DNA methylation, histone modifications, and regulation by non-coding RNAs. All of these epigenetic mechanisms have been proved to affect the main cell types involved in the pathogenesis of SSc, including immune cells, ECs, and fibroblasts (103,152). DNA methylation has been extensively studied in CD4$^+$ T cells from SSc patients, observing co-stimulatory molecules, such as CD70, CD40L and CD11a, to be upregulated in these cell types due to hypomethylation at their promoter regions, as well as a global hypomethylation in IFN-associated genes (153). In the case of SSc dermal fibroblasts, it has been documented that only 6% of differentially methylated CpG sites were shared between dcSSc and lcSSc, observing in both cases a general hypomethylation as compared with controls (153). Histone modifications also play a fundamental role in CD4$^+$ T cells from SSc patients, in which global reduction of H3K27me3, a repressor mark, was observed (154). The first study examining histone modifications in monocytes from SSc patients through chromatin immunoprecipitation (ChIP-seq) and combining it with RNA-seq, showed that genome-wide distribution of H3K4me3 and H3K27ac marks, related with activation of transcription and enhancer regions respectively, were altered in patients as compared with controls (155) Furthermore, variations in these histone marks correlated with genes enriched for immune, IFN and antiviral response pathways, presenting also an overlap with binding sites for transcription factors of the IRF and STAT family. As opposed to CD4$^+$ T cells, elevated H3K27me3 has been reported in fibroblasts from dcSSc patients, generating an overexpression of Fra2, a pro-fibrotic transcription factor (156).

Many of these epigenetic events lead to chromatin conformation modifications, which translate in gene expression changes. These changes can be partially explained by non-coding DNA sequences that act as regulatory elements, which determine where and when genes are turned on or off. Depending on its location and function, regulatory elements include promoters, enhancers, insulators, and silencers (157). The promoter region is located near the start of a gene, serving as a union site for RNA polymerase II and transcription factors, and acting as the starting point for gene transcription. On the other hand, enhancer regions are usually located thousands of bp away from transcription start sites (TSS), acting as binding sites for transcription factors in order to increase transcription level of genes (157).

As previously stated, the majority of SNPs associated with SSc, as well as many other immune-mediated conditions, map to non-coding regions of the genome that are enriched in enhancer elements, which are cell-type specific (130,158). These regulatory elements can interact with genes located hundreds of kilobases away, bypassing nearby genes in many cases, or even located in other chromosomes (159), making spatial chromatin organisation a key mechanism in regulating gene expression.

Genetic studies have the potential to be translated into the clinic (novel drug targets, drug repositioning, personalized medicine, disease and treatment response prediction, etc), but this potential has not been fully realised because of the limitations of GWAS and other high-throughput genotyping studies, that is the identification of causal genes, variants and cell types (160). Thus, the actual challenge remains in linking disease-associated regions with the true genes that are regulating and the specific cell types involved, in order to point to the mechanisms of regulation and the biological pathways implicated in genetically susceptible patients (130). In this regard, many three-dimensional genome architecture techniques have emerged, such as chromosome conformation capture (3C), and subsequent variants (161,162). The most powerful technique developed to date, Hi-C (high-throughput chromosome conformation capture) allows the genome-wide identification of chromosomal interactions within a cell population (163). Nevertheless, the creation of high-resolution maps capable of identifying interactions between regions of interest (such as loci of interest and gene promoters) would need extremely deep sequencing (164). A more recent technique, capture Hi-C (CHi-C) allows to specifically enrich chromosomal regions of interest, such as disease risk loci (region CHi-C) or promoters (promoter CHi-C, pCHi-C) from Hi-C libraries in a cost-effective way (165) (**Figure 8**). This technique has been successfully applied in different cell types to link enhancers and non-coding disease variants to potential target genes (166) as well as to identify disease causal genes and potential drugs for repositioning in ADs using cell lines (167,168) Since the regulation of gene expression is highly context specific, it is essential to apply these technologies to primary cells isolated from patients, to better define the biological mechanisms implicated in disease. In this regard, the largest GWAS performed to date in SSc by our group was implemented with chromatin interaction experiments, concretely through high-resolution maps of enhancer-promoter

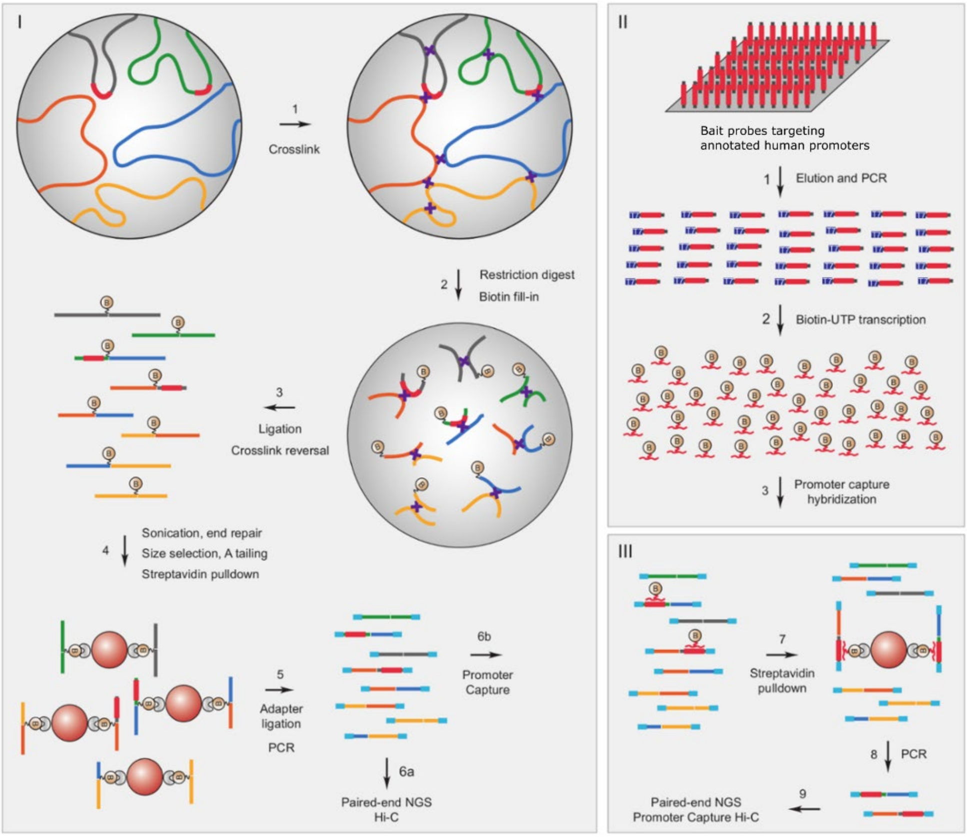


Figure 8. Schematic workflow of promoter capture Hi-C (Modified from Schoenfelder *et al*, J Vis Exp. 2018)

interactions generated by H3K27ac HiChIP in human CD4$^+$ T cells from healthy donors, identifying 43 robust target genes (106). Another study recently published by our group integrated methylation and gene expression data to identify differentially methylated CpG positions and differentially expressed genes in CD4$^+$ T cells from SSc patients and healthy controls, which were confirmed using previously published pCHi-C data, and also combined with SSc GWAS data (169). Thus, we decided to apply pCHi-C technology in two of the most relevant cell types in SSc pathogenesis: CD4$^+$ T cells and CD14$^+$ monocytes primary cells from SSc patients and healthy controls in order to annotate gene targets within known SSc-associated GWAS loci. We also integrated these data with RNA-seq to create a multi-omic approach in order to identify interactomic and transcriptomic differences between cell types and disease states that could be of interest in the pathogenesis of SSc. Results from this study have been consequently included as part of this PhD dissertation.

Genotyping, transcriptomic, and epigenomic studies, taken together, along with characterization studies (CRISPR/Cas9, luciferase assays, etc) comprise a spectrum from early association studies to their functional consequences. The combination of different functional genomics approaches is proving to be useful in the identification of potential causal variants and genes in autoimmunity. In this sense, the whole process, from early discovery of genetic associations through candidate gene or genome-wide association studies to the validation and functional repercussions of these signals, has become a holistic and essential strategy (**Figure 9**). The results of the present PhD dissertation represent an example of this process, helping us to unravel the complex genetic component of SSc.

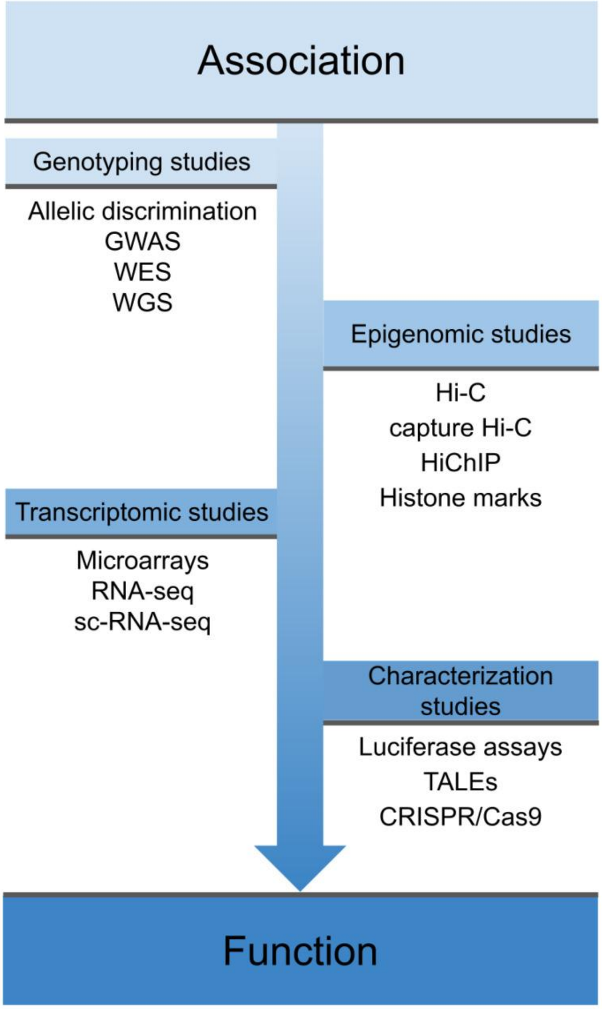


Figure 9. Overview of the spectrum of techniques covered by functional genomics (Extracted from González-Serna *et al*, Genes (Basel). 2020).

OBJECTIVES

The general aim of the present doctoral thesis was to further investigate the genetic component of systemic sclerosis (SSc) to unravel its pathological mechanisms.

The specific objectives were:

1. To identify novel loci associated with the susceptibility to SSc and determine if these loci are associated with the main clinical characteristics

2. To validate previously associated genetic markers in different populations.

3. To further investigate the shared genetic component between SSc and other autoimmune pathologies, such as Crohn's disease or rheumatoid arthritis.

4. To identify differentially expressed genes between SSc patients and controls.

5. To functionally link genetic variants associated with SSc with its gene target through expression quantitative trait locus (eQTLs) analysis and chromosome conformation capture techniques.

MATERIAL AND METHODS, RESULTS, AND DISCUSSION

Chapter 1: Analysis of the genetic component of systemic sclerosis in Iranian and Turkish populations through a genome-wide association study

1.1. Material and methods

1.1.1. Study population

A series of 834 patients diagnosed with SSc (547 from Iran and 287 from Turkey) and 1,455 unaffected and unrelated controls (830 from Iran and 625 from Turkey) were included in this study. All case samples fulfilled the American College of Rheumatology classification criteria for SSc (35,37). Written informed consent was obtained from all the participants. The study was approved by local ethical committees from the different participant centers, in accordance to our institution (Spanish Research Council) ethical committee, and with the Helsinki Declaration of 1975, as revised in 1983.

Clinical information regarding subtypes of SSc, and presence of ACA and ATA were collected. The clinical information was selected based on previous genetic studies determining specific associations with SSc subtypes or autoantibodies status (**Table 1.1**). Some SSc patients showed other complex forms of the disease that could not be classified into dcSSc or lcSSc.

1.1.2. Genotyping

Genomic DNA was extracted from saliva samples or whole blood by standard methods. The GWAS genotyping of the SSc cases and controls was performed as follows: the Iranian and Turkish SSc cases, together with 136 Turkish controls were genotyped using the Illumina Infinium HumanCore-12v1 BeadChip. The remaining Turkish controls were genotyped using the Illumina HumanOmni1-Quad v1 BeadChip. The control group from Iran was genotyped using the Illumina Infinium CoreExome-24 BeadChip.

Table 1.1. Number of subjects and SNPs in the Iranian and Turkish GWASs.

	Number of subjects			
	Iran		Turkey	
	Before QC	After QC	Before QC	After QC
SSc	547	505	287	259
lcSSc	180	165	128	115
dcSSc	301	278	122	111
ACA	30	30	58	47
ATA	384	356	115	102
Healthy Controls	830	770	625	573
	Number of SNPs			
Genotyped	279,616	242,501	236,155	186,435
Imputed	8,467,723	6,313,908	7,369,344	5,885,622
Meta-analysis	5,698,748			

SSc systemic sclerosis, *dcSSc* diffuse cutaneous systemic sclerosis, *lcSSc* limited cutaneous systemic sclerosis, *ACA* anti-centromere antibody, *ATA* anti-topoisomerase antibody, *SNP* single nucleotide polymorphism, *QC* quality control.

1.1.3. Quality control

We applied the same quality control (QC) criteria for both the Iranian and the Turkish GWAS data. SNPs and subjects with call rates lower than 98% and 95%, respectively, were removed using PLINK V.1.9 (170). SNPs with MAFs lower than 0.01 and those that were not in Hardy-Weinberg equilibrium (HWE; *p*-value<0.001) were also excluded. In addition, one subject per duplicate pair and per pair of first-degree relatives was also removed via the Genome function in PLINK V.1.9 with a Pi-HAT threshold of 0.4. Principal component (PC) analyses were performed to identify and exclude outliers based on their ethnicity by using PLINK V.1.9, and the GCTA64 and R-base under GNU Public license V.2. We calculated the 10 first PCs using approximately 100,000 quality-filtered independent SNPs. Those subjects showing more than six standard deviations (SD) from the cluster centroids were considered as outliers. After QC, 764 SSc patients (505 from Iran and 259 from Turkey) and 1,343 controls (770 from Iran and 573 from Turkey) were included in the analysis. A total of 242,501 and 186,435 genotyped SNPs remained after QC filtering of the Iranian and Turkish data sets, respectively (**Table 1.1**).

1.1.4. Imputation of GWAS data

Imputation was performed using the Michigan Imputation Server (MIS) (171). The software SHAPEIT (172) was used in order to estimate the haplotypes, and the European panel in the Haplotype Reference Consortium (HRC) r1.1 (173) was used as reference population for both Turkish and Iranian genotyped data. Imputation was carried out in individual chunks of 50,000 Mb covering whole-genome regions with a probability threshold for merging genotypes of 0.9, to maximize the quality of imputed variants. Imputed data were also subjected to the above-mentioned QC filters in PLINK V.1.9. A total of 6,313,908 and 5,885,622 SNPs were finally analyzed in the Iranian and Turkish GWASs, respectively (**Table 1.1**).

1.1.5. Human leukocyte antigen (HLA) imputation

Considering the previous reported strong HLA association with SSc, a more extensive analysis of the HLA region was conducted in the Turkish and Iranian cohorts. We extracted the extended major histocompatibility complex (MHC) region (29,000,000 to 34,000,000 bp in chromosome 6) from the non-imputed data and imputed a total of 424 classical HLA alleles at two- and four-digits, 7,261 SNPs and 1,276 polymorphic amino acids. We used the SNP2HLA method with the Beagle software package (174) and a reference panel collected by the Type 1 Diabetes Genetics Consortium composed of 5,225 individuals of European origin (175). Alleles and amino acids with call rates <95%, and SNPs showing deviation from HWE (*p*-value<0.001), were removed using PLINK V.1.9.

1.1.6. Statistical analysis

The statistical analyses were performed with PLINK and R. First, the Iranian and Turkish cohorts were independently analyzed by logistic regression on the best-guess genotypes (>0.9 probability) assuming an additive model and including the first 10 PCs as covariates. ORs and 95% confidence intervals (CIs) were calculated according to Woolf's method. Genomic-inflation factor (λ) was estimated in both the Iranian and Turkish cohorts. The value of λ for an equivalent study of 1,000 cases and 1,000 controls ($\lambda_{1,000}$) was also calculated. We also performed a stratified analysis considering the different clinical and serological features (dcSSc, lcSSc, ACA+SSc and ATA+SSc). Subsequently, the Iranian and the Turkish GWAS data sets were combined by inverse variance weighted fixed effects meta-analysis to integrate the two independent association studies. Cochran's Q and I^2 tests were used to measure the heterogeneity of the ORs across studies. SNPs showing association *p*-values $<5.0\times10^{-8}$ were regarded as significant, and those showing *p*-values $< 1.0\times10^{-5}$ were regarded as suggestive associations. The presence of independent effects was examined using a stepwise logistic regression by conditioning on the lead SNP and the first 10 PCs.

Regarding the HLA region, to determine the influence of the polymorphic amino acid positions on disease susceptibility, an omnibus association test was carried out in the Iranian and Turkish cohorts as described (176). A null generalized linear model was established, including the 10 first PCs as covariates. Next, an alternative model, including the same variables and all the possible alleles in the analyzed amino acid positions was built. Finally, to compare both models, a likelihood ratio test was conducted.

1.2. Results

Following QC and imputation, a total of 5,698,748 SNPs in 764 SSc patients (505 from Iran and 259 from Turkey) and 1,343 controls (770 from Iran and 573 from Turkey) were analyzed. The final λ showed minimal evidence of population stratification in the Iranian and Turkish cohorts (λ=1.02 in both cases). Additionally, we calculated $\lambda_{1,000}$ with consistent results in the Iranian ($\lambda_{1,000}$=1.03) and Turkish ($\lambda_{1,000}$=1.04) cohort.

As shown in **Figure 1.1**, a high association peak corresponding to the HLA class II region reached the GWAS significance threshold (5x10^{-8}) in both the Iranian and Turkish cohorts. Specifically, for the Iranian cohort, the top associated signal belonged to the SNP rs9268923 (*p*-value=6.55x10^{-18}, OR=2.29) (**Figure 1.1A**). Out of the HLA region, some suggestive associations were found, highlighting the *IRF5-TNPO3* (interferon regulatory factor 5-transportin 3) and the *NFKB1* (Nuclear Factor Kappa B Subunit 1) loci (**Table 1.2**), which represent previously reported susceptibility loci for SSc. The top associated signal for the *IRF5-TNPO3* locus corresponded to the SNP rs17424921 (*p*-value =5.28x10^{-5}, OR=1.64) located in an intergenic region and 5' upstream from the *TNPO3* gene. In our dataset, this SNP is in strong LD (r^2=0.92) with rs10488631 (*p*-value=6.15x10^{-5}, OR=1.62), a SNP previously associated to SSc (91,177) that is located in the intergenic region between *IRF5* and *TNPO3*. Regarding the *NFKB1* locus, the top associated SNP was the intronic variant rs4648133 (*p*-value=1.52x10^{-5}, OR=1.48). Interestingly, this SNP is in moderate LD (r^2=0.73) with rs1598859 (*p*-value=6.23x10^{-3}, OR=1.27), previously associated to SSc and located in an intronic position of the *NFKB1* gene (96).

In the Turkish cohort (**Figure 1.1B**), the top associated SNP within the HLA class II region was rs34297496 (*p*-value =9.46x10^{-9}, OR=2.16). Out of the HLA region, one signal corresponding to the SNP rs7095491 reached the suggestive level of significance (*p*-value =5.17x10^{-7}, OR=1.88). This SNP is located in an intergenic region between *GOT1* (glutamic-oxaloacetic transaminase 1) and *NKX2.3* (NK2 Homeobox 3). This locus was specifically associated in the Turkish cohort and did not reach the suggestive level of significance in either the Iranian cohort (*p*-value=3.14x10^{-2}, OR=0.89) or the meta-analysis (*p*-value=3.21x10^{-1}, OR=1.07).

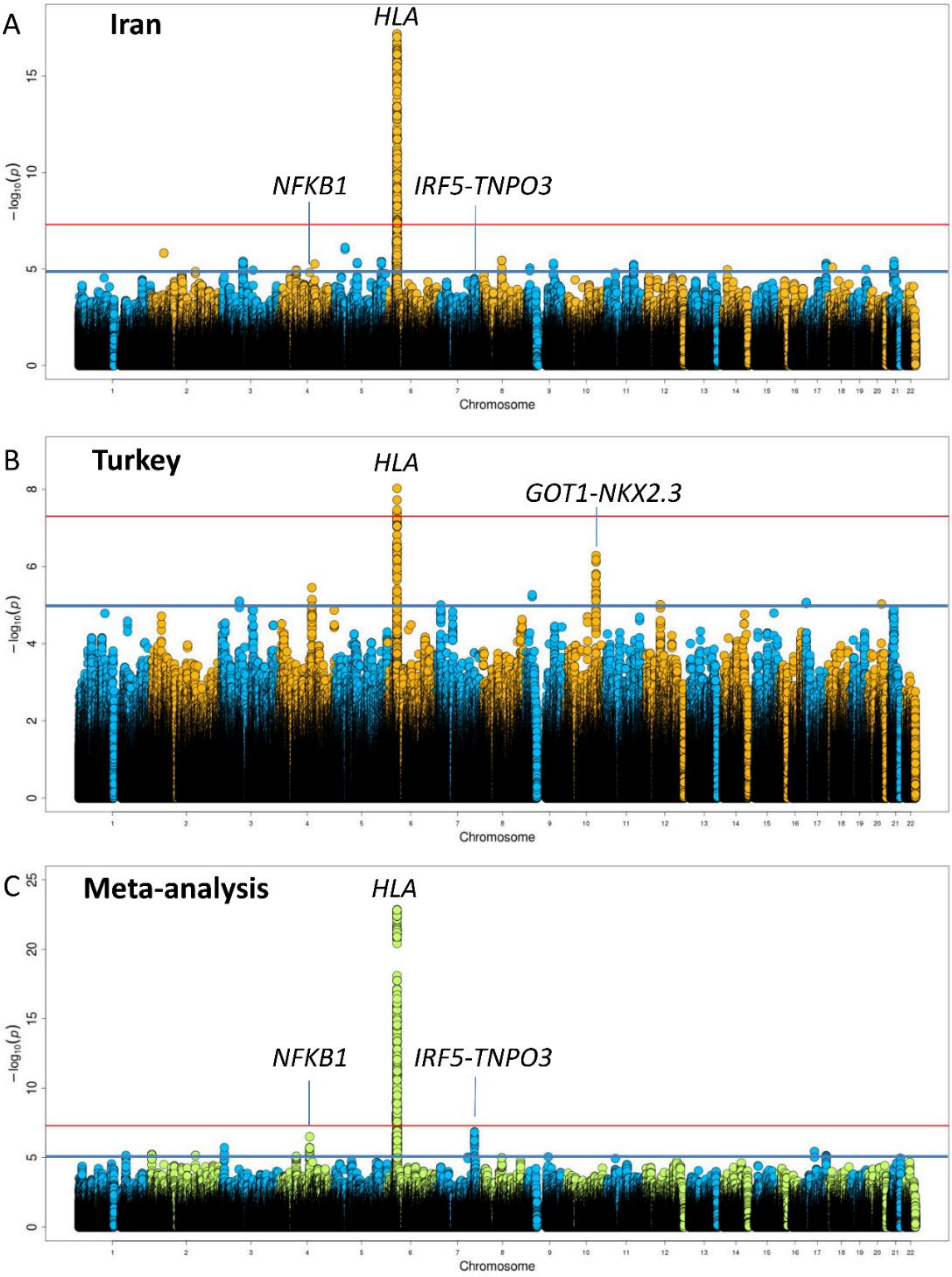


Figure 1.1. Manhattan plot of the GWAS results from the Iranian cohort (A), Turkish cohort (B) and meta-analysis (C). The values on the y-axes denote the -log10 transformed *p*-values. Genomic position for each SNP for 22 autosomes are plotted on the x-axis. The red line denotes the a priori threshold for genome-wide significance (*p* -value= 5x10^{-8}). The suggestive level of significance (*p*-value = 1x10^{-5}) is highlighted in blue.

Table 1.2. Non-HLA loci associated with SSc at the suggestive significance level (p-value<1x10^{-5}) in the meta-analysis.

					Iran		Turkey		Meta-analysis			
Loci	**Chr**	**BP***	**Most significant SNP**	**Minor allele**	***p*-value**	**OR (95% CI)**	***p*-value**	**OR (95% CI)**	***p*-value**	**OR (95% CI)**	**Q**	**I^2**
IRF5-TNPO3	7	128708122	rs17424921	C	5.28E-05	1.64 (1.29-2.08)	6.57E-04	1.77 (1.27-2.46)	1.34E-07	1.68 (1.39-2.03)	0.69	0
	7	128594183	rs10488631	C	6.15E-05	1.62 (1.28-2.05)	3.05E-03	1.66 (1.18-2.33)	6.35E-07	1.63 (1.34-1.97)	0.90	0
NFKB1	4	103536413	rs4648133	C	1.52E-05	1.48 (1.24-1.77)	6.29E-03	1.45 (1.11-1.89)	3.11E-07	1.47 (1.27-1.70)	0.90	0
GRM7-LOC101927394	3	7907077	rs9821717	T	2.52E-05	0.68 (0.57-0.82)	2.23E-02	0.74 (0.57-0.96)	1.91E-06	0.70 (0.60-0.80)	0.59	0
UNC45B	17	33501644	rs12452554	C	3.77E-04	1.74 (1.28-2.35)	2.54E-03	2.00 (1.28-3.13)	3.57E-06	1.81 (1.41-2.33)	0.61	0
PXDN-MYT1L	2	1785659	rs2059413	C	2.29E-04	0.71 (0.60-0.85)	8.01E-03	0.70 (0.54-0.91)	5.70E-06	0.71 (0.61-0.82)	0.92	0
LOC101929282-RBM43	2	151773679	rs917238	G	1.73E-05	1.54 (1.27-1.88)	9.15E-02	1.28 (0.96-1.72)	6.86E-06	1.45 (1.23-1.71)	0.30	6.49
ATF6-OLFML2B	1	161937892	rs3002626	A	7.41E-04	0.73 (0.60-0.87)	2.38E-03	0.65 (0.50-0.86)	6.98E-06	0.70 (0.60-0.82)	0.52	0
HOPX-SPINK2	4	57558089	rs9968446	A	1.21E-05	1.44 (1.22-1.70)	1.46E-01	1.21 (0.93-1.56)	7.92E-06	1.37 (1.19-1.57)	0.25	22.08
OSTF1-PCSK5	9	78042927	rs473299	A	1.42E-04	1.44 (1.19-1.74)	2.02E-02	1.38 (1.05-1.80)	8.69E-06	1.42 (1.21-1.65)	0.78	0

*BP corresponding to the NCBI build 37.

BP base pair, *SNP* single nucleotide polymorphism, *OR* odds ratio, *CI* confidence interval, *Q and I*^{2} heterogeneity values.

1.2.1. Meta-analysis

Subsequently, we decided to perform a meta-analysis in order to identify susceptibility loci to SSc with moderate effect sizes in the Iranian and the Turkish cohorts, independently. The HLA class II region was the highest association peak observed in the meta-analysis (**Figure 1.1C**), and the top associated signal belonged to the SNP rs28746976 (*p*-value=1.41x10^{-23}, OR=2.09). Out of the HLA region, two signals corresponding to the *IRF5-TNPO3* and *NFKB1* loci almost reached the genome-wide significance threshold (**Table 1.2**). The top associated signals corresponded to the same SNPs observed in the Iranian cohort for both loci, *IRF5-TNPO3* rs17424921 (*p*-value=1.34x10^{-7}, OR=1.68) and *NFKB1* rs4648133 (*p*-value=3.11x10^{-7}, OR=1.47) (**Figure 1.2**).

1.2.2. HLA analysis

In order to extensively analyze the association of the HLA region with SSc in both the Iranian and the Turkish populations, a comprehensive HLA imputation was performed. As stated above, a strong association signal was observed within the HLA class II region in both the Iranian and the Turkish GWAS data sets (**Table 1.3**). Regarding the Iranian cohort, the top associated signal belonged to the classical allele *HLA-DRB1*11:04* (*p*-value=2.10x10^{-24}, OR=3.14). After controlling for *HLA-DRB1*11:04*, an independent secondary effect was found in the HLA-DPB1 region. Specifically, the top associated signal was the *HLA-DPB1*13:01* allele (*p*-value=5.37x10^{-14}, OR=5.75). No additional independent associations were observed after conditioning on both signals, *DRB1*11:04* and *DPB1*13:01* (**Figure 1.3**). Significant insights were found in the ATA+SSc patients vs. controls stratified analysis, as a stronger association between the *HLA-DRB1*11:04* and the disease was evident (*p*-value=2.49x10^{-34}, OR=4.92). In addition, after controlling for this classical allele, the *HLA-DPB1*13:01* also showed a stronger association (*p*-

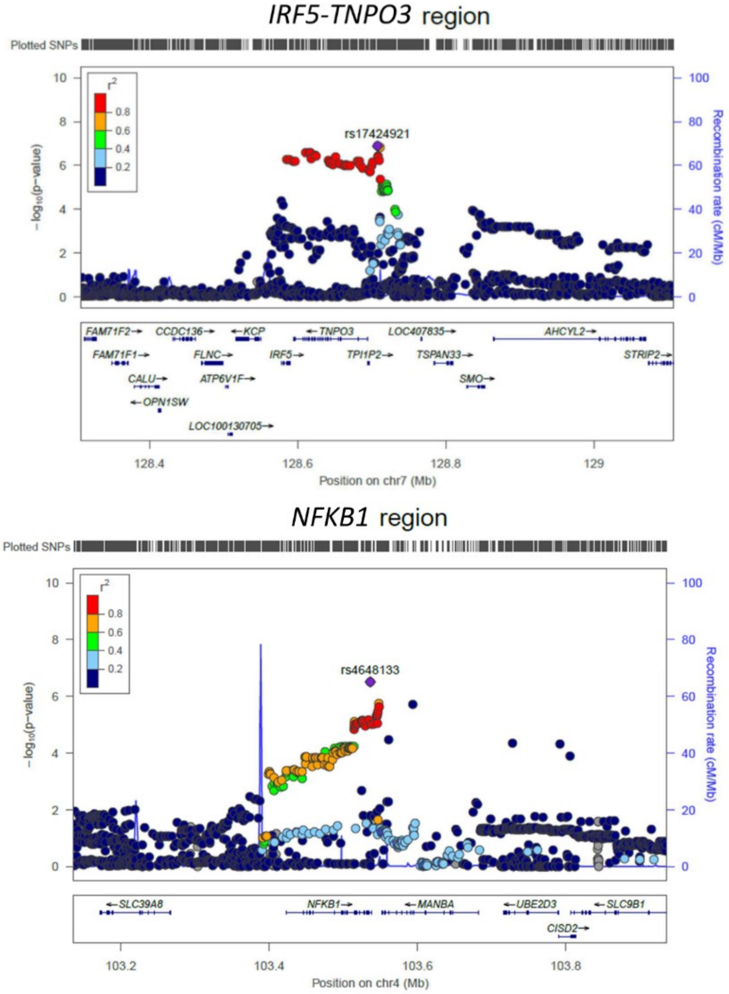


Figure 1.2. Locus zoom plot for *IRF5-TNPO3* and *NFKB1* regions. The *p*-values for association (−log10 values) of each SNP are plotted against their physical position on chromosome 7 and 4 for *IRF5-TNPO3* and *NFKB1* regions, respectively.

value=6.44x10^{-21}, OR=10.60). We also performed a stratified analysis comparing ATA+SSc vs. ATA-SSc, showing a strong association of the *HLA-DRB1*11:04* with ATA+SSc (*p*-value= 2.69x10^{-14}, OR=5.06). After conditioning on *HLA-DRB1*11:04*, the *HLA-DPB1*13:01* allele also showed statistically significant association with ATA+SSc (*p*-value= 5.02x10^{-8}, OR=10.21). In the

Turkish cohort, *DRB1*11:04* (*p*-value=4.90x10^{-11}, OR=2.93) showed the most significant association, and no independent secondary effects were found after controlling for this HLA classical allele (**Figure 1.4**).

Subsequently, specific amino acid positions that could be responsible for the association observed for these classical alleles were examined by means of an omnibus test. The most relevant amino acid positions for disease risk in the Iranian cohort were the positions 58 (P_{LRT} =6.24x10^{-24}) and 67 (P_{LRT} =5.76x10^{-22}) of the HLA-DRβ1 molecule (**Table 1.4**), which were in strong LD (r^2=0.83). After conditioning on the strongest association (position 58), position 76 (P_{LRT} =1.57x10^{-13}) of the HLA-DPβ1 protein remained independently associated. None of the other signals remained significant after conditioning on positions 58 and 76 of the HLA-DRβ1 and HLA-DPβ1 molecules, respectively. The most associated amino acid residues for the previously mentioned positions were Glu (*p*-value=5.39x10^{-22}, OR=2.57) and Phe (*p*-value=1.08x10^{-20}, OR=2.38) in positions 58 and 67 of the HLA-DRβ1 molecule, respectively, and Ile (*p*-value=1.52x10^{-12}, OR=4.60) in position 76 of the HLA-DPβ1. In the case of the Turkish cohort, the position 58 of the HLA-DRβ1 molecule was the only signal reaching the significance threshold (P_{LRT} =4.96x10^{-8}) (**Table 1.3**), being Ala the most associated amino acid residue in that position (*p*-value=9.42x10^{-8}, OR=2.13).

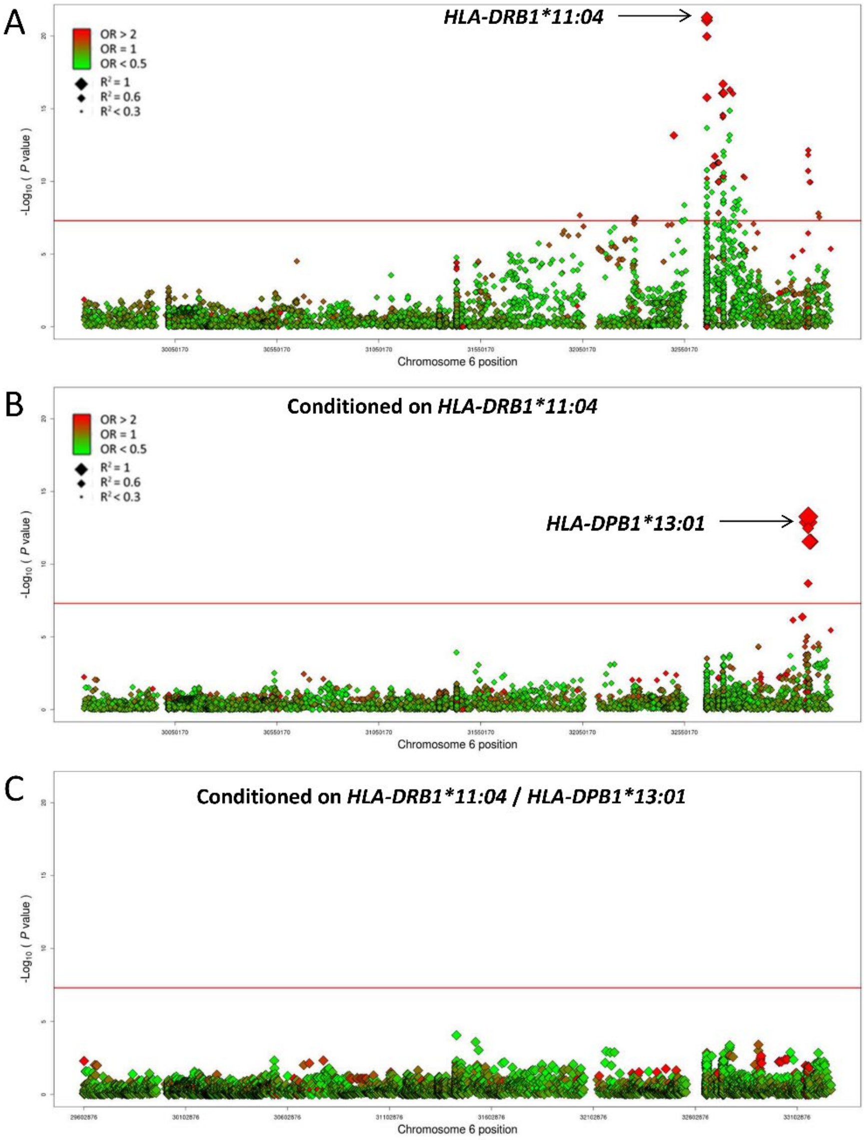


Figure 1.3. Manhattan plot representation of the results of the conditional logistic regression analysis of the HLA region in Iranian patients with SSc. (A) Unconditioned test of the HLA region. (B) Results after conditioning on the *HLA-DRB1*11:04* classical allele. (C) Results after conditioning on the *HLA-DRB1*11:04* and *DPB1*13:01* alleles. The red/green color gradient represents the effect direction of each analyzed variant (red for risk and green for protection). The size of the diamonds indicates the degree of linkage disequilibrium with the classical allele *HLA-DRB1*11:04* and *DPB1*13:01* in A and B, respectively. The red line represents the genome-wide level of significance (*p*-value = $5x10^{-8}$).

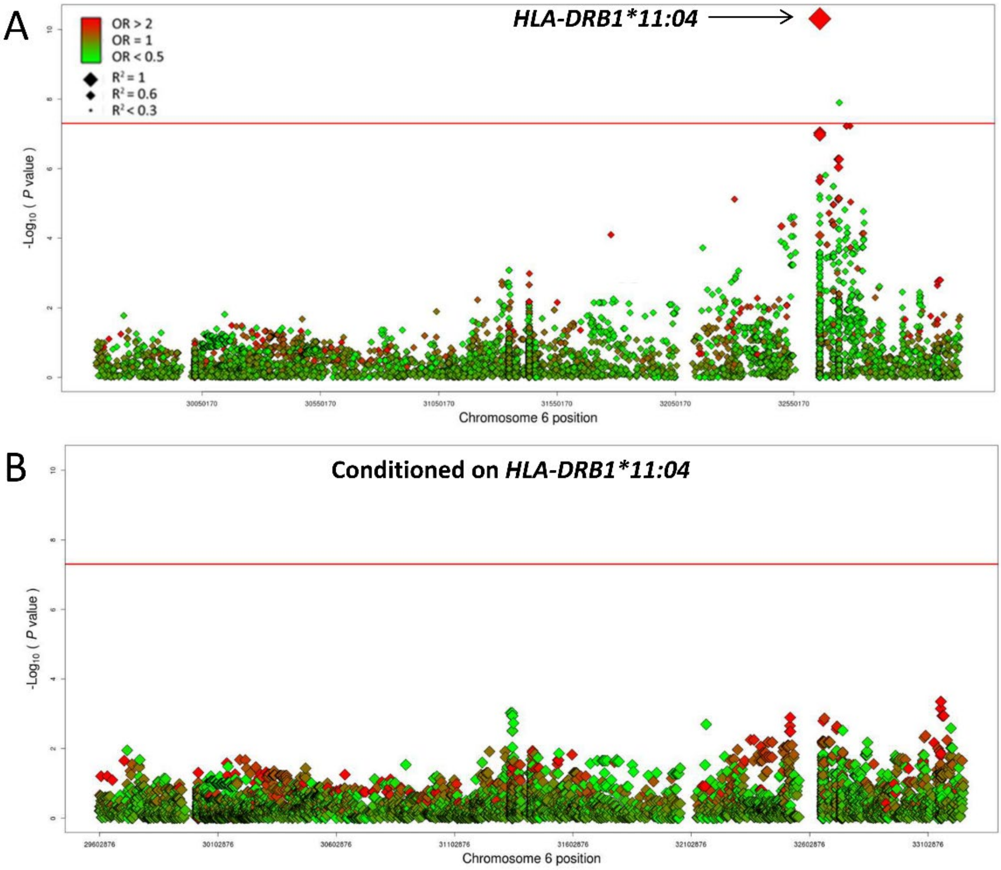


Figure 1.4. Manhattan plot representation of the results of the conditional logistic regression analysis of the HLA region in Turkish patients with SSc. (A) Unconditioned test of the HLA region. (B) Results after conditioning on the *HLA-DRB1*11:04* classical allele. The red/green color gradient represents the effect direction of each analyzed variant (red for risk and green for protection). The red line represents the genome-wide level of significance (p-value = 5x10^{-8}).

Table 1.3. Classical four digit HLA alleles showing the strongest association with SSc.

	Iran				Turkey			
	Unconditioned		Conditioned on HLA-DRB1*11:04		Unconditioned		Conditioned on HLA-DRB1*11:04	
HLA allele	***p*-value**	**OR (95% CI)**	***p*-value**	**OR (95% CI)**	***p*-value**	**OR (95% CI)**	***p*-value**	**OR (95% CI)**
DRB1*11:04	**2.10E-24**	**3.14 (2.52-3.91)**	NA	NA	**4.90E-11**	**2.93 (2.12-4.03)**	NA	NA
DQB1*03:01	2.04E-17	2.22 (1.84-2.66)	3.08E-02	1.31 (1.03-1.68)	9.16E-07	1.93 (1.49-2.51)	4.51E-01	1.15 (0.80-1.64)
DPB1*13:01	**7.23E-13**	**5.05 (3.24-7.86)**	**5.37E-14**	**5.75 (3.65-9.07)**	1.79E-03	2.93 (1.49-5.75)	4.48E-04	3.43 (1.72-6.84)
DQA1*05:01	1.08E-10	1.77 (1.49-2.11)	6.54E-01	1.05 (0.84-1.31)	1.08E-05	1.77 (1.37-2.28)	3.46E-01	1.16 (0.85-1.59)

HLA human leucocyte antigen, *OR* odds ratio, *CI* confidence interval.

Table 1.4. Amino acid positions showing the strongest association with SSc after the omnibus test.

				Iran		Turkey	
HLA molecule	**Amino acid position**	**Center codon position**	**Tested Alelles**	***P_{LRT}* (unconditioned)**	***P_{LRT}* conditioned on DRβ1 position 58**	***P_{LRT}* (unconditioned)**	***P_{LRT}* conditioned on DRβ1 position 58**
DRβ1	**58**	**32659974**	**2**	**6.24E-24**	NA	**4.96E-08**	NA
DRβ1	67	32659947	3	5.76E-22	4.18E-02	6.19E-06	8.40E-01
DQβ1	45	32740702	2	1.14E-17	4.67E-01	3.68E-07	4.15E-01
DQβ1	55	32740672	3	1.25E-16	1.59E-02	6.95E-03	2.72E-01
DQβ1	140	32737868	2	2.47E-16	3.68E-02	3.97E-03	4.05E-01
DQβ1	182	32737742	2	2.47E-16	3.68E-02	3.97E-03	4.05E-01
DPβ1	**76**	**33156640**	**3**	**8.06E-14**	**1.57E-13**	8.40E-03	3.93E-03

HLA human leucocyte antigen, *LRT* likelihood ratio test.

1.3. Discussion

As previously stated in the Introduction section, the study of the genetic background of diseases in different non-Caucasian population in order to break the existing European bias for risk prediction represents a major challenge. In this regard, this study represents the first SSc GWAS performed in Iranian and Turkish populations. Given the relevance of the HLA region in SSc predisposition and the lack of information about the role of HLA genes in SSc in Iranian or Turkish populations, an extensive analysis of the HLA region was performed. The HLA class II region was identified as the most strongly associated *locus* to SSc in both Middle Eastern populations. Interestingly, the *HLA-DRB1*11:04* classical allele showed the most significant association in both ancestries. These data reinforce the role of the *HLA-DRB1*11:04* allele in the SSc susceptibility previously reported in different ethnic populations, including Caucasians (125,178,179), African-Americans (125) and Mexicans (126). Furthermore, these results suggest for the first time that this effect could be driven by the amino acid Glu-58. Nevertheless, amino acid Phe-67, which is in high LD with Glu-58 in our study, has been previously set as a relevant amino acid in SSc susceptibility (101). There is no functional implication described in the literature for these two amino acids so we cannot assure which one is leading the association. In addition, after conditioning on *HLA-DRB1*11:04*, the classical allele *HLA-DPB1*13:01*, which has been reported as a susceptibility allele in Caucasians (125) and Koreans (90), remains significant in the Iranian cohort. Our results also suggest that the association between *HLA-DPB1*13:01* and SSc could be driven by amino acid Ile-76, which is in line with previous results (101). Interestingly, amino acid position 76 is part of the binding pocket of the HLA-DPβ1 molecule (180). It should be noted that, in previously mentioned studies, associations of *HLA-DRB1*11:04* and *HLA-DPB1*13:01* with SSc were explained by its strong correlation with ATA. This specific correlation with ATA+SSc patients was verified in our stratified analysis. In this sense, the association of *HLA-DRB1*11:04* and *HLA-DPB1*13:01* classical alleles with the global disease could be owing to a clinical predominance of ATA+SSc patients in the Iranian population. On the other hand, the lack of association of *HLA-DPB1*13:01* with the ATA+SSc subgroup in the Turkish cohort could be due to its lower sample size as compared with the Iranian cohort. Due to the low allele frequency of *HLA-DPB1*13:01* in our Turkish samples (4%), small changes in the sample size can result in huge changes in the statistical power; e.g., considering an OR= 3.2 and an allele frequency of 4%, the statistical power in the Turkish cohort was 25% whereas in the Iranian cohort was 85%.

Outside of the HLA region, two suggestive associations at the *IRF5-TNPO3* and *NFKB1* loci were found. *IRF5-TNPO3* was one of the first risk loci identified in SSc. This association has been replicated in various studies and ethnicities (91,106,181,182), indicating that it is a firm susceptibility factor for SSc and other ADs such as SLE and RA (183,184). On the other hand, the *NFKB1* locus was identified as a SSc susceptibility gene in a previous GWAS meta-analysis and subsequently confirmed in a candidate gene approach analysis in a Caucasian population (96,185). Nevertheless, until the posterior release of the largest meta-GWAS in SSc performed to date (106), our study represented the strongest association described for this locus, being performed in a much smaller sample size, which emphasize the importance of performing large-scale genotyping studies in different populations in order to discover new associated loci. NF-κB has been broadly described as controlling the inflammatory process, and its role in autoimmunity is widely accepted (186). Furthermore, the interaction of *NFKB1* with other well-defined susceptibility genes in SSc, such as *TNFAIP3* (Tumor Necrosis Factor Inducible Protein A20) (187), which encodes a protein that inhibits NF-κB activation, suggests that it could be a good candidate gene to be involved in SSc. Regarding the Turkish cohort, a suggestive level associated signal corresponding to the *GOT1-NKX2.3* locus emerged. This locus is a well-established signal associated with other immune-mediated diseases, such as ulcerative colitis and CD (188,189), suggesting a potential role of the *GOT1-NKX2.3* locus in autoimmunity. Notably, the strongest association reported for the *GOT1-NKX2.3* locus in both studies (rs4409764) is in strong LD (r^2=0.93) with the top associated SNP observed in our Turkish cohort (rs7095491). *NKX2.3* (Nirenberg-Kim (NG) 2 homeobox 3) is a homeodomain transcription factor essential for the correct development of spleen and small intestine (190,191). Interestingly, this gene is expressed in microvascular endothelial cells, and its overexpression has been associated with both CD and ulcerative colitis through truncated regulation of VEGF signaling and the production of endothelin-1 (190,191). These processes are intimately related with SSc vasculopathy, which highlights *NKX2.3* as a good candidate gene contributing to SSc pathogenesis.

Despite the successful identification of SSc susceptibility genes in Iranian and Turkish populations, our study had some limitations. In this regard, larger cohorts could help to elucidate whether the two suggestive associations observed in the meta-analysis (*IRF5-TNPO3* and *NFKB1*) reach the genome-wide significance threshold. Nevertheless, the results for four of the six most relevant SSc non-HLA associations reported by Carmona *et al* (192) in a Turkish population were very similar to our GWAS results (**Table 1.5**), highlighting the reproducibility of our study.

In summary, our results confirm the previously reported association of the HLA region with SSc susceptibility and show two non-HLA associations almost reaching the genome-wide significance threshold, in Iranian and Turkish populations. This study sheds light on the unexplored genetic background of SSc in these populations, which contributes to a better understanding of the genetic structure and pathogenesis of the disease.

Table 1.5. Comparative of the results obtained in Turkish population in four SSc hits. In this table are shown the results obtained by Carmona *et al* (192) and our Turkish GWAS.

SNP	Gene	Study	*p*-value	OR
rs10488631	*IRF5-TNPO3*	Turkish GWAS	3.05E-03	1.66
		Carmona *et al*	1.32E-05	1.76
rs3821236	*STAT4*	Turkish GWAS	3.60E-02	1.34
		Carmona *et al*	6.50E-02	1.21
rs9373839	*ATG5*	Turkish GWAS	8.86E-01	1.02
		Carmona *et al*	6.72E-01	1.06
rs2056626	*CD247*	Turkish GWAS	6.93E-03	0.71
		Carmona *et al*	2.20E-03	0.75

SNP single nucleotide polymorphism, *OR* odds ratio, *GWAS* genome-wide association study.

Chapter 2: Analysis of the shared genetic component between systemic sclerosis and Crohn's disease through a cross-disease meta-GWAS

2.1. Material and methods

2.1.1. Study population

A series of 5,734 patients diagnosed with SSc, 4,588 CD patients, and 14,568 healthy controls of European origin were enrolled in this study. **Figure 2.1** and **Supplementary Table S2.1** detail the cohorts included in the different stages of the study.

SSc GWAS dataset: In the discovery phase, we included GWAS data from 2,281 SSc cases and 4,410 healthy controls from Spain, USA, Germany and the Netherlands, all of them included in a previous study (91).

CD GWAS dataset: The CD discovery cohort was composed of 1,988 cases and 2,978 healthy controls from the UK, included in the CD GWAS performed by the Welcome Trust Case Control Consortium (WTCCC) (193).

Replication cohorts: To confirm the results obtained in the discovery phase, genotyping data of the selected polymorphisms were obtained from GWAS data from 3,453 SSc cases and 3,602 controls, and 2,600 CD cases and 3,578 controls. Specifically, the SSc replication cohort included three independent case/control sets from Spain, USA, and Italy. Regarding the CD cohort, case/control sets were recruited from Spain, USA and Germany, all of them from previously published GWASs (194–196).

The control population consisted of unrelated healthy individuals that were recruited in the same geographical regions as patients. Genotyping information of each cohort is included in **Supplementary Table S2.1**.

All SSc cases were defined based on the 1980 preliminary and 2013 classification criteria of American College of Rheumatology (35,37) or based on the presence of at least 3 out of 5 CREST features typical for SSc. All CD cases were defined based on a confirmed diagnosis of CD using conventional endoscopic, radiological and histopathological criteria (197).

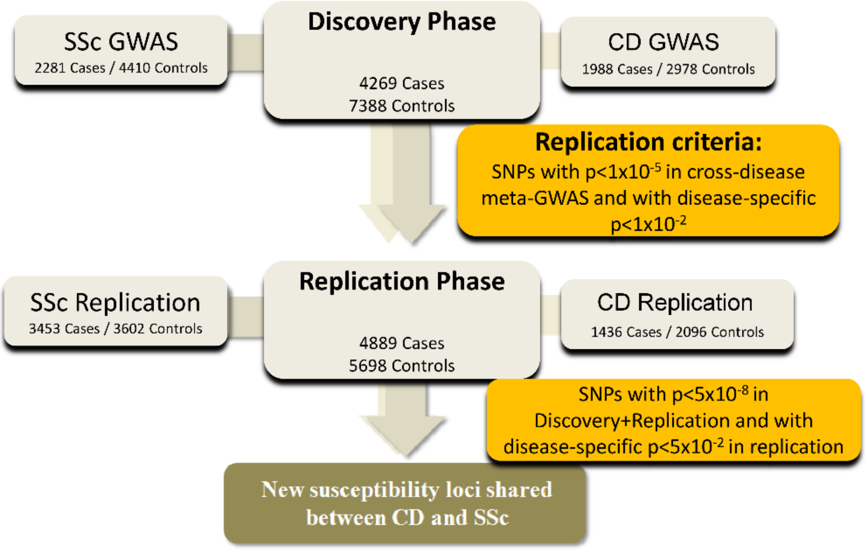

SSc GWAS
2281 Cases / 4410 Controls
Discovery Phase
4269 Cases
7388 Controls
CD GWAS
1988 Cases / 2978 Controls
Replication criteria:
SNPs with p<1x10-5 in cross-disease meta-GWAS and with disease-specific p<1x10-2
SSc Replication
3453 Cases / 3602 Controls
Replication Phase
4889 Cases
5698 Controls
CD Replication
1436 Cases / 2096 Controls
SNPs with p<5x10-8 in Discovery+Replication and with disease-specific p<5x10-2 in replication
New susceptibility loci shared between CD and SSc

Figure 2.1. Schema of the study design.

2.1.2. Quality control and imputation

All GWAS data were QC filtered prior imputation. SNPs and subjects with success call rates lower than 95% were removed using PLINK V.1.9 (170). SNPs showing a deviation from the Hardy–Weinberg equilibrium (*p*-value<0.001) and minor allele frequencies <1% were also excluded. In addition, one subject per duplicate pair and per pair of first-degree relatives was also removed via the Genome function in PLINK V.1.9 with a Pi-HAT threshold of 0.4. Principal component analysis (PCA) was performed in order to identify and exclude outliers based on their ethnicity by using PLINK V.1.9 and the GCTA64 and R-base under GNU Public license V.2. We estimated the first five PCs using ~100.000 quality-filtered independent SNPs (r^2<0.15).

Outliers were defined as individuals who deviated more than six standard deviations from the centroid of their population. The number of SNPs before and after QC for each cohort is summarized in **Supplementary Table S2.1**.

Imputation was performed using the Michigan imputation Server (171). The software SHAPEIT (172) was used in order to estimate haplotypes, and the European panel of the HRC r1.1 (173) was used as the reference panel for both SSc and CD genotype data in the discovery phase. Individual chunks of 50.000 Mb were used to carry out the imputation, covering whole-genome regions with a probability threshold for merging genotypes of 0.9, thus maximizing the quality of the imputed variants. Imputed data were also subjected to the above-mentioned QC filters in PLINK V.1.9. The total number of SNPs imputed for each cohort is summarized in **Supplementary Table S2.1**.

2.1.3. Statistical analysis

Statistical analyses were performed with PLINK V.1.9.

Discovery phase: Each GWAS case/control cohort was independently analyzed by logistic regression assuming an additive model with the first five PCs as covariates, as a correcting method for population stratification. ORs and 95% CIs were calculated according to Woolf's method. Subsequently, SSc datasets were meta-analysed by the inverse variance-weighted method. Sex chromosomes were excluded from the analysis.

In order to detect common signals for SSc and CD with the same effect, either risk or protection, summary statistics of each disease were then meta-analyzed applying the inverse variance method. To identify common signals for SSc and CD with opposite effect, the direction of association was flipped in the CD dataset (1/OR instead of OR) before the SSc-CD meta-analysis. Subsequently, we selected SNPs that showed a p-value < 1 x 10-5 in the SSc-

CD meta-analysis and showed nominal significance (p-value < 0.01) with each disease separately, as well as no significant heterogeneity in the SSc meta-analysis (Cochran's Q test > 0.05 and heterogeneity index I2 < 50%).

The strongest associated SNP within each locus was selected for the replication phase. Genetic variants were annotated using variant effect predictor (VEP) (198) and their previous association with SSc and/or CD was explored using Immunobase (http://www.immunobase.org) and the GWAS catalog (199).

Replication phase: Replication cohorts were analysed by logistic regression for the previously selected SNPs. Finally, combined analysis of the SSc and CD discovery and replication cohorts was performed using the inverse variance method. After the replication phase, we considered as statistically significant those signals that showed a *p*-value < 0.05 in each disease separately in the replication phase and a *p*-value < 5×10^{-8} in the SSc-CD cross-disease meta-analysis including both discovery and replication datasets.

The statistical power of the SSc-CD combined meta-analyses (both discovery and discovery+replication) was determined as described by Skol *et al* (200).

Independence analysis: For those SSc-CD common loci identified, for which an association with any of the analysed diseases was already reported, we evaluated the independence between pleiotropic signals and genetic variants previously associated with SSc and/or CD at the genome-wide significance level according to Immunobase and the GWAS Catalog. For this purpose, we used LDlink (201), a tool that provides LD data between polymorphisms across a variety of ancestral populations. Only the European ancestry was taken into account for the LD analysis.

In addition, since one of the shared genetic risk loci was located close to the extended MHC region, we decided to test the independence between our new common signal and the main SSc and CD HLA associations. For this, we imputed SNPs, classical HLA alleles and amino acids across the extended MHC region (29,000,000 to 34,000,000 bp in chromosome 6) using the SNP2HLA method with the Beagle software package (174) and the Type 1 Diabetes Genetics Consortium reference panel (175). HLA imputation of the CD discovery cohort was not possible due to the low coverage of this region included in the platform used for the genotyping of this dataset. For the SSc discovery cohort, the presence of independent effects within the extended MHC region was examined using a stepwise logistic regression by conditioning on the top independent signals.

2.1.4. Functional annotation

We assessed the potential regulatory function of the SSc-CD common susceptibility variants identified by means of *in silico* eQTL analysis using Haploreg v4.1. Haploreg v4.1 is a tool for exploring annotations at variants on haplotype blocks, providing a large collection of regulatory information, capable of the functional assignment onto any set of variants derived from GWAS or sequencing studies (202). We only included eQTLs found in tissues with relevance in SSc and/or CD.

2.1.5. Protein-protein interaction and gene set enrichment analysis

In order to identify interactions among proteins encoded by the SSc and CD common risk loci, we decided to construct a protein-protein interaction (PPI) network using the STRING database V.11.0 (203). This software provides a critical assessment and integration of PPI, including functional (indirect) as well as physical (direct) associations. The interaction confidence score was set in 0.9, which is the highest score calculated as a combined probability from different evidences of interactions corrected for the probability of observing an interaction by random chance.

In addition, gene ontology (GO) was applied to perform an enrichment analysis in order to determine whether certain biological processes are overrepresented in the set of SSc-CD common genes.

2.2. Results

2.2.1. Meta-analysis and replication

Following QC and imputation, we performed a meta-analysis considering both diseases as a single phenotype. A total of 5,994,231 SNPs overlapped between all GWAS datasets in the discovery phase. In the discovery cross-disease meta-analysis, the statistical power to detect an association at a *p*-value of $1x10^{-5}$ (MAF=20% and OR=1.2) was 80%. In the discovery+replication meta-analysis, the statistical power to detect an association at a *p*-value of $5x10^{-8}$ (MAF=20% and OR=1.2) was 100%.

When we combined GWAS data from SSc and CD under the assumption that alleles had the same effect in both diseases, genetic variants at 13 loci fulfilled the replication criteria (*p*-value < $1x10^{-5}$ in the SSc-CD meta-GWAS and *p*-value < 0.01 in each disease-specific analysis) (**Figure 2.2A** and **Supplementary Table S2.2**). One of these common signals was located within the *IRF8* region, a known genetic risk locus shared between SSc and CD, and, therefore, it was not considered in subsequent analyses. On the other hand, we performed the analysis under the assumption that alleles had opposite directions in both diseases, identifying 12 loci that fulfilled all criteria for the replication phase (**Figure 2.2B** and **Supplementary Table S2.3**).

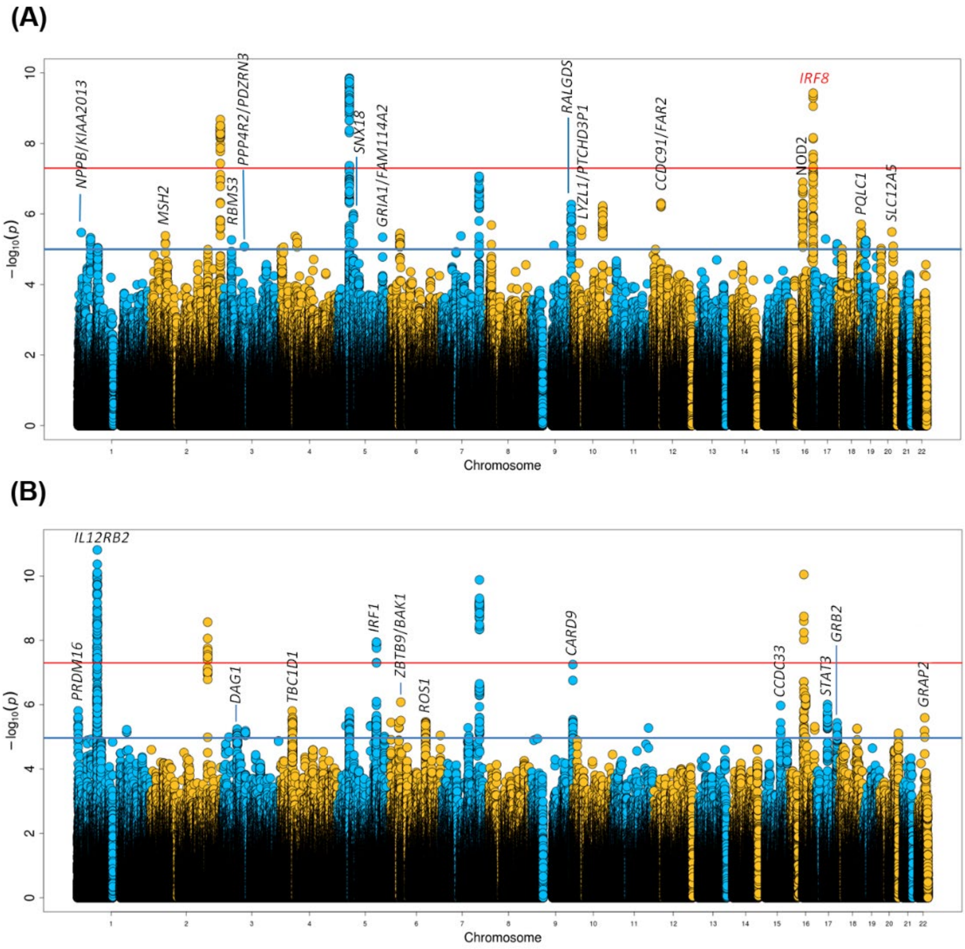


Figure 2.2. Manhattan plot representing the results of the cross-disease meta-analysis including systemic sclerosis and Crohn's disease, considering same allelic effects (A) and opposite allelic effects (B). Loci selected for replication are marked in black. Significance threshold at the genome-wide significance level is marked with a red line. Established significance threshold for the cross-disease meta-analysis (*p*-value < 1x10^{-5}) is marked with a blue line.

To confirm these associations, the strongest associated SNP within each locus was selected for validation in additional sample sets. According to the criteria established for the replication analysis (genome-wide significance in the combined analysis including both discovery and replication sets, and nominal statistical significance in each disease-specific replication analysis), we identified a total of 4 genetic variants showing a pleiotropic effect in SSc and CD: two intronic variants located within *IL12RB2* and *STAT*3, a SNP close to *IRF1*, and an intergenic variant at 6p21.31 located between *ZBTB9* and *BAK1* (**Table 2.1**). It is remarkable that an opposite allelic effect in both disorders was observed for all these new common signals.

Three of these shared risk loci have been previously associated with one of the analysed diseases, *IL12RB2* with SSc and *IRF1* and *STAT3* with CD. Shared genetic variants at the *IRF1* and *STAT3* loci identified in our study were linked to those polymorphisms previously associated with CD (r^2> 0.40). In the case of *IL12RB2*, it is an established genetic risk locus for SSc but, in addition, the *IL23R* gene, located within this same genomic region, is a known susceptibility gene for CD. However, LD analysis evidenced that the pleiotropic variant identified in our study (rs6659932) was independent of the *IL23R* SNPs previously associated with CD (**Supplementary Table S2.4**).

On the other hand, the intergenic variant at 6p21.31 (rs68191) is located close to the extended MHC region. Considering this, we decided to test the independence between our new common signal and the main HLA associations observed in the SSc and CD discovery cohorts. In the case of CD, independence between signals could not be checked due to the low coverage of the HLA region. Regarding SSc, two independent signals were observed after conditional regression analysis, *HLA-DPB1*1301* (p-value=1.77x10^{-19}, OR=2.79) and *HLA-DRB1*1104* (p-value=1.21x10^{-12}, OR=1.83). After controlling for these two classical alleles, the SSc-CD common signal remained significant in the SSc discovery cohort (*p*-value=8.15x10^{-3}; conditioned *p*-value=2.78x10^{-2}).

Table 2.1. Loci associated at the genome-wide significance threshold after the cross-disease meta-analysis of systemic sclerosis and Crohn's disease. Results of the discovery and replication analysis for each individual disease and the combined meta-analysis (discovery + replication) are shown.

				Discovery					**Replication**				**Discovery+ Replication**
				SSc		**CD**		**SSc-CD**	**SSc**		**CD**		**SSc-CD**
Region	**Gene**	**SNP**	**Test Allele**	***p*-value**	**OR**	***p*-value**	**OR**	***p*-value**	***p*-value**	**OR**	***p*-value**	**OR**	***p*-value**
1p31.3	*IL12RB2*	rs6659932	A	2.47E-08	1.3	1.33E-04	0.79	1.54E-11	3.75E-03	1.13	3.44E-02	0.86	1.08E-11
5q31.1	*IRF1*	rs2548998	G	1.55E-03	1.27	3.09E-07	0.79	1.13E-08	2.00E-02	1.08	1.18E-03	0.88	2.18E-11
6p21.31	*ZBTB9/ BAK1*	rs68191	C	8.15E-03	0.84	8.70E-06	1.39	8.33E-07	2.09E-04	0.82	1.56E-02	1.15	1.07E-10
17q21.2	*STAT3*	rs4796791	T	1.34E-03	1.13	1.52E-04	0.84	9.86E-07	3.85E-02	1.08	1.85E-02	0.90	2.52E-08

SNP single nucleotide polymorphism, *SSc* systemic sclerosis, *CD* Crohn's disease, *OR* odds ratio.

2.2.2. Functional effect on gene expression

Subsequently, we used the HaploReg database to explor wether the most strogly associated polymorphism of each shared locus acted as an eQTL. As shown in **Supplementary Table S2.5**, all the pleiotopic SNPs identified in our study appeared to affect gene expression levels. Shared genetic variants at the *IL12RB2* (rs6659932) and *STAT3* (rs4796791) loci affected expression levels of *IL12RB2* and *STAT3*, respectively, whereas the pleiotropic SNP of the *IRF1* locus (rs2548998) acted as an eQTL for *IRF1* and *SLC22A5*. Interestingly, the intergenic polymorphism at the MHC extended region (rs68191) affected gene expression levels of *TAPBP*.

2.2.3. Protein-protein interaction and enrichment analysis

Finally, we also evaluated the connectivity at the protein interaction level among the genetic risk loci shared between SSc and CD, including genes whose expression levels were affected by the pleiotopic polymorphisms identified in our study, that is *IRF1*, *SLC22A5*, *STAT3*, *IL12RB2* and *TAPBP*, as well as loci associated in previous studies with both SSc and CD, including *STAT4*, *TYK2*, *IRF8*, *GSDMA* and *IKZF3*. *GSDMA* and *IKZF3* belong to the same LD block, however *GSDMA* has been set as the most probable candidate gene of this locus in SSc and *IKZF3* for CD (127,188). Thus, we decided to keep both genes for PPI and enrichment analyses.

The PPI network involved 9 of the 10 common proteins included in the analysis, except for SLC22A5 (**Figure 2.3**). We observed a strongly significant PPI enrichment (*p*-value < 1x10^{-6}), indicating that these proteins have more interactions than would be expected for a random set of proteins of similar size.

To further evaluate this connection, we performed a gene ontology enrichment analysis in biological processes. In this regard, we observed 29 statistically significant over-represented biological processes (*p*-value < 0.05). The most significantly over-represented pathways were related to interleukin-mediated signaling, especially those related with the IL-12 family and the type I IFN signaling pathway (**Table 2.2**).

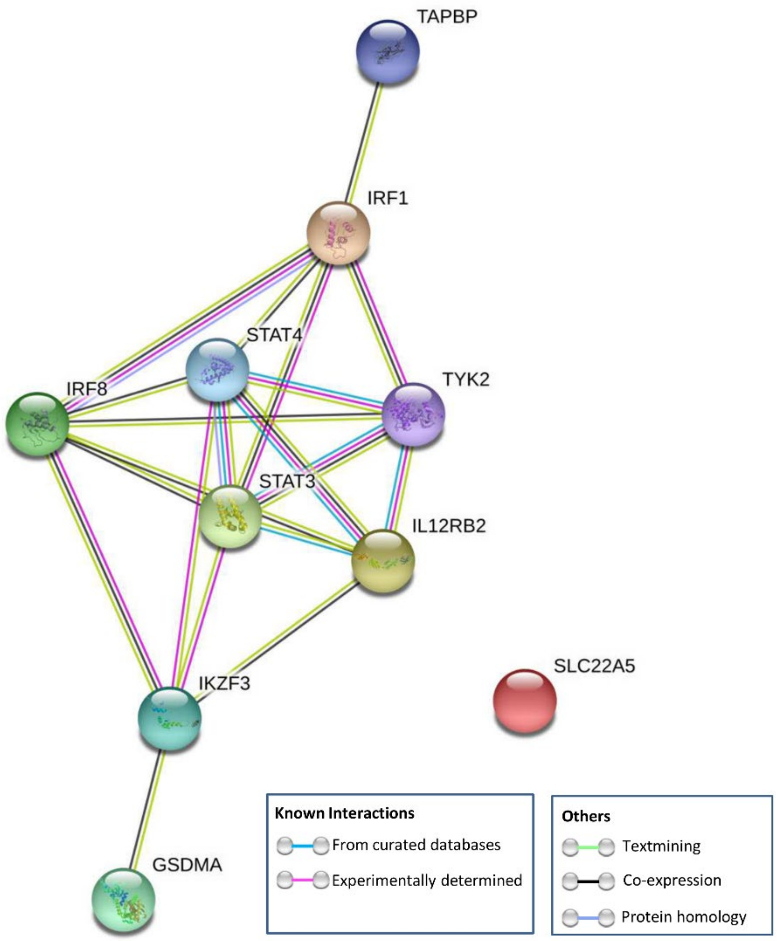


Figure 2.3. STRING protein-protein interaction network connectivity among genetic risk loci shared between systemic sclerosis and Crohn's disease

Table 2.2. Most significantly enriched Gene Ontology (GO)-biological processes in the set of genetic risk loci shared between systemic sclerosis and Crohn's disease.

Biological pathway	**GO term**	***p*-value***	**Count in gene set**	**Shared genes involved**
Interleukin-35-mediated signaling pathway	GO:0070757	1.44E-05	3 of 11	*STAT4*, ***STAT3***, ***IL12RB2***
Interleukin-23-mediated signaling pathway	GO:0038155	1.44E-05	3 of 9	*STAT4*, ***STAT3***, *TYK2*
Cytokine-mediated signaling pathway	GO:0019221	6.16E-05	6 of 655	*STAT4*, ***STAT3***, ***IL12RB2***, *TYK2*, ***IRF1***, *IRF8*
Interleukin-12-mediated signaling pathway	GO:0035722	3.20E-04	3 of 47	*STAT4*, ***IL12RB2***, *TYK2*
Type I interferon signaling pathway	GO:0060337	3.60E-04	3 of 65	*TYK2*, ***IRF1***, *IRF8*
Interleukin-21-mediated signaling pathway	GO:0038114	6.00E-04	2 of 8	*STAT4*, ***STAT3***
Interleukin-27-mediated signaling pathway	GO:0070106	8.30E-04	2 of 11	***STAT3***, *TYK2*
Positive regulation of transcription by RNA polymerase II	GO:0045944	4.90E-03	5 of 1104	*STAT4*, ***STAT3***, ***IRF1***, *IRF8*, *IKZF3*
Positive regulation of interleukin-12 production	GO:0032735	5.90E-03	2 of 34	***IRF1***, *IRF8*
Receptor signaling pathway via JAK-STAT	GO:0007259	7.90E-03	2 of 41	*STAT4*, ***STAT3***
Alpha-beta T cell differentiation	GO:0046632	9.70E-03	2 of 50	***STAT3***, ***IRF1***

**p*-values determined by binomial statistic test and adjusted by false discovery rate correction.
New loci shared between systemic sclerosis and Crohn's disease are in bold.

2.3. Discussion

Through the first comprehensive study of the genetic component shared between SSc and CD, we have identified four loci that contribute to suceptibility to both disorders. Of these, one had not been previously associated with any of the diseases under study (an intergenic locus at 6p21.31), whereas the remaining three represent established genetic risk loci for one but not the other condition.

Although all these pleiotropic SNPs are located in non-coding regions, functional annotation indicated that they act as regulatory variants affecting expression levels of either the gene where they mapped or close genes in cell types or tissues of relevance in the pathogenesis of SSc and/or CD. In this regard, pleiotropic variants appeared to influence expression levels of the *IL12RB2*, *IRF1*, *SLC22A5*, *STAT3*, and *TAPBP* genes (**Supplementary Table S2.5**) Most of these genes are key players of the immune response: *IL12RB2* encodes a subunit of the IL-12 receptor complex implicated in Th1 differentiation; *STAT3* encodes a transcription factor that is essential for the differentiation of Th17 cells; *IRF1* encodes a transcriptional regulator of type I IFN and IFN-inducible genes; and *TAPBP* is crucial for optimal peptide loading on the MHC class I molecule. In addition, the pleiotropic variant affecting *IRF1* levels also regulates the expression of *SLC22A5*, which encodes an organic cation transporter involved in the active cellular uptake of carnitine.

Interestingly, PPI analysis evidenced a number of non-random connections among the SSc-CD common genes, including both shared risk loci previously described and comon genes identified in our study, which indicates overlap among the pathways involved in the pathogenesis of these two disorders. Specifically, the IL-12 family signaling pathways, including IL-35, IL-23, IL-12, IL-21, and IL-27-mediated signaling, were particularly compelling. This family of cytokines plays a crucial role in shaping immune responses, differentiation of naive T cells towards different types of effector cells, as well as in the regulation of effector cell functions (204) (**Figure 2.4**). In this sense, *IRF1*, *STAT3,* and *IL12RB2* play a particularly interesting role in Th1/Th17 regulation. Moreover, the type I IFN signaling pathway was also enriched among the set of SSc-CD common genes. As previously stated, increased expression and activation of IFN-inducible genes, known as interferon signature, has been extensively reported in SSc (205) and several IRFs, including *IRF5*, *IRF4*, and *IRF8*, have been involved in its susceptibility (84,106), thus supporting the role of *IRF1*, previously associated with CD but not with SSc, as a new susceptibility gene for this last condition.

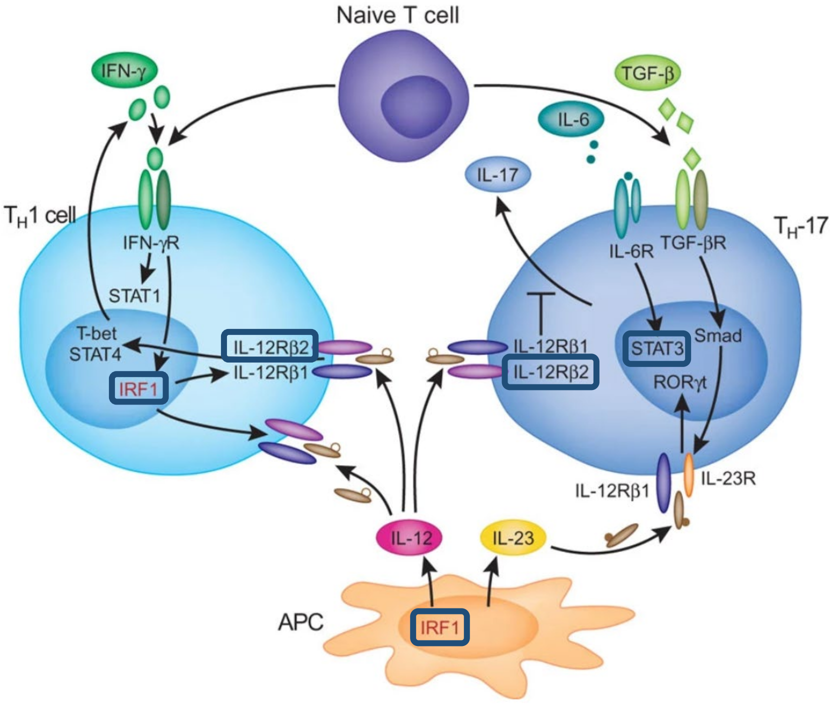


Figure 2.4. Implication of the new shared associated genes in IL-12 pathway and naive T cell differentiation (Modified from Unutmaz *et al*, Nat Immunol. 2018)

Considering these results, both IL-12 family and type I IFN signaling pathways could represent interesting therapeutic targets for both SSc and CD. Indeed, ustekinumab, a monoclonal antibody to the p40 subunit common to IL-12 and IL-23, has been recently approved in the EU and the USA to treat patients with CD and, therefore, this drug could be repositioned to treat SSc. However, it should be advised that all the pleiotropic variants identified in our study showed opposite allelic effects in the two analysed disorders, thus highlighting the complex effects that shared associations have on disease outcomes. This could be due to the fact that consequences of genetic variants are influenced by the cell type. For example, as previously indicated, the shared genetic variant at *IL12RB2* influenced its gene expression levels; however, whereas the minor allele (which conferred risk to SSc in our study) correlated with an increased gene expression in whole blood, the major allele (which conferred risk to CD) had the same effect (increased *IL12RB2* expression) in fibroblasts, according to Genotype-Tissue Expression project (GTEx) data (206). In addition, the effect on gene expression of the pleiotropic SNP located within the 5q31.1 region was also cell type specific, influencing *IRF1* expression levels in lymphoblastoid cells and *SLC22A5* levels in other tissues, and, therefore, this SNP could have a different biological implication in both diseases. Indeed, higher expression levels of OCTN2, the protein encoded by *SLC22A5*, have been found in inflamed regions of the intestinal epithelium compared with non-inflamed areas, and a role of this protein in the intestinal homeostasis has also been reported (207); whereas, given the relevance of the type 1 IFN signaling pathway in SSc, the *IRF1* gene seems a more plausible candidate to be involved in SSc susceptibility. Considering this, it is possible that an effective treatment for SSc could have a detrimental effect on CD, and conversely. As previously mentioned, we observed discordant associations for variants located in genes implicated in IL-23 and Th1 differentiation pathways. In this context, IL-17-specific antibody therapy, effective in psoriasis and with promising effects on SSc (208,209), has been proven to exacerbate CD (210). This could be due to a deficient Th17 activation in CD owing to mutations in *STAT3*, which could lead to hyper-IgE syndrome, typically associated with extracellular fungal and bacterial infections (211). Interestingly, according to our results, the *STAT3* rs4796791 variant confers protection to CD and risk to SSc, which could lead to an exacerbate reaction in CD patients carrying this variant when treated with anti-IL17 therapy.

Interestingly, it has been reported a reduced incidence of CD in patients with SSc (212,213). Although the causes of this phenomenon are not clear, our results suggest that identical genetic risk factors could have different or even opposite functional effects in both diseases. These 'flip-flop' associations have been extensively observed across different comparative analyses (214). In this regard, a cross-disease meta-analysis including CD and type 1 diabetes (215) identified two variants, *IL27* rs4788084 and *IL10* rs3024505, with opposite effects in these two conditions. Furthermore, a meta-analysis of 6 different immune-mediated disorders showed that 14% of overlapped variants were discordant regarding the risk allele across diseases (216). These results suggest that predisposition to related diseases may be regulated by different dose balance of genes and genomic elements in relevant biological pathways, as well as how these differences affect a specific cell type, as previously mentioned. In this sense, differences across cell types in transcription regulation mediated by epigenetic factors, such as methylation or histone modifications, as well as derived changes in chromatine conformation, could influence these opposite effects for the same allele in different diseases (217). It is, therefore, crucial to know the cell types in which genetic variants are acting to be able to elucidate their role on the pathogenesis of the disease.

2.4. Supplementary Data

Supplementary Table S2.1. Study cohorts.

		N				Genotyping platform				Genotyped SNPs		Imputed SNPs	
		SSc	**Controls**	**CD**	**Controls**	**SSc**	**Controls**	**CD**	**Controls**	**Before QC**	**After QC**	**Before QC**	**After QC**
Discovery	**USA**	1482	2759	-	-	Illumina HumanHap550 K	Breast and prostate cancer controls CGEMS; Illumina iControlDB	-	-	477,663	461,068	15,556,150	7,347,081
	Spain	362	362	-	-	Illumina HumanCNV370 K	Illumina HumanCNV370K	-	-	330,55	317,348	10,739,039	6,885,748
	Germany	255	658	-	-	Illumina HumanCNV370 K	Illumina HumanHap550k	-	-	299,278	290,928	10,236,621	7,052,424
	Netherlands	182	631	-	-	Illumina HumanCNV370 K	Illumina HumanHap550k	-	-	290,564	286,266	10,749,243	7,143,940
	UK	-	-	1988	2978	.	-	Affymetrix GeneChip 500K	Affymetrix GeneChip 500K	380,935	379,593	11,633,951	6,854,680
Replication	**USA**	1286	1388	956	982	Illumina HumanCore	HumanHap300v1.1	Illumina HumanHa p300k	Illumina HumanHap 300k	-	-	-	-
	Spain	1169	1262	1164	1482	Illumina HumanCore; HumanCytoSNP -12v2	Illumina HumanCore	Illumina Quad610 Beadchip	Illumina Quad610 Beadchip	-	-	-	-
	Italy	998	952	-	-	Illumina HumanCore	Illumina HumanHap550k	-	-	-	-	-	-
	Germany	-	-	480	1114	-	-	Illumina HumanHa p550k	Illumina HumanHap 550k	-	-	-	-

N number, SSc systemic sclerosis, *CD* Crohn's disease, *QC* quality control.

Supplementary Table S2.2. Genetic variants that reached the replication criteria when considering same allelic effect in systemic sclerosis and Crohn's disease. Results of each disease-specific meta-analysis and the cross-disease meta-analysis for the strongest associated genetic variant within each locus are shown.

		VEP annotation				**SSc**		**CD**		**Cross-disease meta-GWAS**		
Region	**SNP**	**Consequence**	**Mapped gene**	**Candidate gene**	**Test Allele**	***p*-value**	**OR**	***p*-value**	**OR**	***p*-value**	**OR**	**Previously reported**
16q24.1	rs11642873	Intergenic	-	*IRF8*	C	3.78E-07	0.77	1.99E-04	0.81	3.73E-10	0.79	SSc, CD
16q12.1	rs72798422	Intergenic	-	*NOD2*	C	1.76E-02	1.26	1.23E-07	1.78	1.26E-07	1.47	CD
12p11.22	rs6487699	Intergenic	-	*CCDC91/FAR2*	C	3.45E-05	1.17	3.92E-03	1.14	4.92E-07	1.16	-
9q34.2	rs694881	Intronic	*RALGDS*	*RALGDS*	G	2.13E-04	0.86	7.07E-04	0.84	5.34E-07	0.85	-
5q11.2	rs2059214	Intronic	*SNX18*	*SNX18*	T	3.73E-05	1.19	7.07E-03	1.14	1.00E-06	1.17	-
18q23	rs72984220	Intronic	*PQLC1*	*PQLC1*	G	2.47E-03	1.17	1.46E-04	1.27	1.96E-06	1.21	-
10p11.23	rs11598403	Intergenic	-	*LYZL1/PTCHD3P1*	A	3.37E-03	0.83	1.39E-04	0.75	2.83E-06	0.8	-
20q13.12	rs2297199	Intronic	*SLC12A5*	*SLC12A5*	C	3.09E-03	1.14	7.66E-04	1.19	3.23E-06	1.14	CD
1p36.22	rs198382	Intergenic	-	*NPPB/KIAA2013*	C	5.32E-03	1.21	9.41E-05	1.37	3.35E-06	1.27	-
2p16.3	rs4953504	Intronic	*MSH2*	*MSH2*	C	6.78E-03	0.87	8.87E-05	0.79	4.13E-06	0.84	-
5q33.2	rs2614119	Intergenic	-	*GRIA1/FAM114A2*	G	5.10E-03	0.9	1.55E-04	0.85	4.54E-06	0.88	-
3p24.1	rs1347772	Intronic	*RBMS3*	*RBMS3*	T	1.98E-04	0.87	8.22E-03	0.88	5.41E-06	0.87	-
3p13	rs6765560	Intergenic	-	*PPP4R2/PDZRN3*	C	9.78E-04	1.13	2.66E-03	1.14	8.42E-06	1.14	-

VEP variant effect predictor, *SNP* single nucleotide polymorphism, *SSc* systemic sclerosis, *CD* Crohn's disease, *OR* odds ratio.

Supplementary Table S2.3. Genetic variants that reached the replication criteria when considering opposite allelic effect in systemic sclerosis and Crohn's disease. Results of each disease-specific meta-analysis and the cross-disease meta-analysis for the strongest associated genetic variant within each locus are shown.

		VEP annotation				SSc		CD		Cross-disease meta-GWAS		
Region	**SNP**	**Consequence**	**Mapped gene**	**Candidate gene**	**Test Allele**	***p*-value**	**OR**	***p*-value**	**Inverted OR**	***p*-value**	**OR**	**Previously reported**
1p31.3	rs6659932	Intronic	*IL12RB2*	*IL12RB2*	A	2.47E-08	1.3	1.33E-04	1.26	1.54E-11	1.28	SSc
5q31.1	rs2548998	Intronic	*AC116366.3*	*IRF1*	G	1.55E-03	1.27	3.09E-07	1.26	1.13E-08	1.18	CD
9q34.3	rs3812565	Upstream	*CARD9*	*CARD9*	C	1.86E-03	0.88	2.70E-06	0.81	5.67E-08	0.85	CD
6p21.31	rs68191	Intergenic	-	*ZBTB9/BAK1*	C	8.15E-03	0.84	8.70E-06	0.72	8.33E-07	0.79	-
17q21.2	rs4796791	Intronic	*STAT3*	*STAT3*	T	1.34E-03	1.13	1.52E-04	1.19	9.86E-07	1.16	CD
15q24.1	rs12905224	Intronic	*CCDC33*	*CCDC33*	C	2.15E-04	1.16	1.47E-03	1.16	1.06E-06	1.15	-
1p36.32	rs2742661	Intronic	*PRDM16*	*PRDM16*	T	6.29E-04	1.2	6.53E-04	1.25	1.56E-06	1.22	-
4p14	rs17578878	Intronic	*TBC1D1*	*TBC1D1*	T	9.23E-05	1.26	5.02E-03	1.23	1.57E-06	1.25	-
22q13.1	rs138014	Downstream	*GRAP2*	*GRAP2*	T	3.11E-04	0.87	2.53E-03	0.87	2.54E-06	0.87	-
6q22.1	rs72969416	Intronic	*ROS1*	*ROS1*	C	2.67E-04	0.8	3.91E-03	0.82	3.43E-06	0.81	-
17q25.1	rs4350602	Intronic	*GRB2*	*GRB2*	C	4.91E-04	0.86	2.27E-03	0.85	3.69E-06	0.86	-
3p21.31	rs4625	Downstream	*DAG1*	*DAG1*	G	6.82E-03	0.9	1.36E-04	0.83	5.72E-06	0.87	CD

VEP variant effect predictor, *SNP* single nucleotide polymorphism, *SSc* systemic sclerosis; *CD* Crohn's disease, *OR* odds ratio.

Supplementary Table S2.4. Linkage disequilibrium between the systemic sclerosis-Crohn's disease common signal at *IL12RB2* (rs6659932) and the *IL23R* polymorphisms previously associated with Crohn's disease.

		LD with rs6659932		
SNP	**Position**	**r^2**	**D'**	**Study accession number**
rs7517847	1:67215986	0.0051	0.1372	GCST003044
rs11581607	1:67242007	0.0059	0.1328	GCST004132
rs11465804	1:67236843	0.0124	0.1854	GCST000207
rs11209026	1:67240275	0.0059	0.1328	GCST001396
rs11805303	1:67209833	0.0043	0.2282	GCST000042
rs76418789	1:67182913	0.0004	1	GCST002094

SNP single nucleotide polymorphism, *LD* linkage disequilibrium.

Supplementary Table S2.5. Potential role of the lead pleiotropic polymorphisms as expression quantitative trait loci (eQTLs) in tissues of relevance for the diseases under study.

Region	Locus	SNP	Correlated gene	Tissue	*p*-value	Study
5q31.1	*IRF1*	rs2548998	*SLC22A5*	Whole Blood	2.95E-07	[1]
				Esophagus - Mucosa	5.09E-06	[1]
				Cells - Transformed fibroblasts	6.82E-06	[1]
				Lymphoblastoid cells	4.94E-07	[2]
			IRF1	Lymphoblastoid cells	1.33E-06	[2]
1p31.3	*IL12RB2*	rs6659932	*IL12RB2*	Cells - Transformed fibroblasts	3.41E-13	[1]
				Whole Blood	5.47E-07	[1]
17q21.2	*STAT3*	rs1026916*	*STAT3*	Whole Blood	1.99E-20	[3]
6p21.31	*ZBTB9/BAK1*	rs68191	*TAPBP*	Whole Blood	4.04E-06	[3]

*Proxy SNP of the strongest associated polymorphism, rs47967941 (r^2=0.99).

SNP single nucleotide polymorphism.

[1] GTEx Consortium. Science. 2015;348(6235):648-60.

[2] Lappalainen T, *et al.* Nature. 2013;501(7468):506-11

[3] Westra HJ, *et al.* Nat Genet. 2013;45(10):1238-43.

Chapter 3: Genome-wide meta-analysis study of the shared genetic component in systemic seropositive rheumatic diseases

3.1. Material and methods

3.1.1. Study population

This study was conducted using 12,132 affected subjects and 23,260 controls of European descent population. All of them were included in previously published GWAS as summarized in **Supplementary Table S3.1** (91,218–222). Briefly, a total of 3,255 SLE cases and 9,562 ancestry matched controls were included from six countries across Europe and North America (Spain, Germany, Netherlands, Italy, UK, and USA). All of the included patients were diagnosed based on the standard American College of Rheumatology (ACR) classification criteria (223). Previously described GWAS data from 2,363 SSc cases and 5,181 ancestry matched controls were included in the study (four case-control collections from Spain, Germany, Netherlands and USA). All the patients met the ACR Preliminary criteria for the classification of SSc or had at least 3 of the 5 CREST features (35,37). A total of 4,804 RA cases and 3,793 ancestry matched controls were included from Sweden, UK and USA, obtained from the Epidemiological Investigation of RA (EIRA) project (http://www.eirasweden.se), the WTCCC data repositories, (http://www.wtccc.org.uk/), and the North American Rheumatoid Arthritis Consortium (NARAC), respectively. All the patients met the ACR criteria for the diagnosis of RA (224) or were diagnosed by board-certified rheumatologists. IIM GWAS data were obtained in collaboration with the MYOGEN consortium, comprising 1,710 cases and 4,724 ancestry matched controls from Europe and North America (Spain, Sweden/Netherlands, Czech Republic/Hungarian, USA and UK). The inclusion criteria were defined by proximal weakness, myopathy on electromyography, muscle biopsy consistent with idiopathic inflammatory myopathy or elevated serum muscle enzymes, and the presence of Gottron's papules/sign or heliotrope rash, with exclusion of other causes of muscle disease per Bohan and Peter criteria (225).

3.1.2. Quality control and imputation

Unified QC of the 18 case-control collections was conducted separately, based on stringent criteria using PLINK v.1.9 (170). Given that related and/or duplicated subjects may have been recruited for different studies, genome-wide relatedness was assessed and one individual from each pair was removed. Samples with <95% of successfully called genotypes were removed.

Further, SNPs with genotyping call rate <98%, MAF <1% and deviating from HWE with a *p*-value <0.001 in the control group were removed. To control for possible population stratification, we performed PC analysis using GCTA64 and R-base software under GNU Public license v.2. For that, the first ten PCs were calculated for each individual and those samples found >6 standard deviations from the cluster centroids of each set were considered outliers and were removed from the analyses.

Imputation of autosomal SNPs was conducted in the Michigan Imputation Server using Minimac3 (171). The software SHAPEIT (172) was used for haplotype reconstruction and the HRC r1.1 (173) was used as the reference population.

3.1.3. Statistical analysis

Disease-specific association testing: Association testing for allele dosages was performed using EPACTS software (https://genome.sph.umich.edu/wiki/EPACTS) adjusting by the first two or five PCs as appropriate. Additionally, prior to the meta-analysis, each individual study was adjusted by their specific inflation factor. This was performed by multiplying each standard error by the square root of the calculated inflation factor.

Cross-disease meta-analysis: meta-analysis was conducted with METASOFT (226). Fixed-effects or random-effects model was applied depending on study heterogeneity (Cochran's Q test *p*-value). The lead and most significant SNP outside the extended MHC region, was selected to perform sequential conditional association analyses to confirm statistical independence of associations. These analyses were carried out with the software GCTA-COJO (227,228), using the summary statistics from the SSc, SLE, RA and IIM meta-analysis, with the GCTA "--cojo-cond" option within a 10 Mb window. A secondary signal was considered if within this window there were additional SNPs with a conditioned *p*-value passing a Bonferroni correction.

Variant annotation: To annotate the association signals we utilized SNPnexus (229). If a SNP mapped in a coding/non-coding gene region (introns, exons, UTRs), we reported the candidate gene identified by SNPnexus. If the SNP was upstream, downstream, or intergenic, the most "likely" gene was manually curated in a 500kb window and assigned based on the observed LD block and if there were evidences of previous association signals with ADs or other immune-related phenotypes.

Model search to identify the diseases contributing to the lead signal: we conducted an exhaustive disease-subtype model search with the R statistical software package ASSET (230), utilizing each case-control collection separately. This subset-based meta-analysis explores all possible subsets of diseases for the presence of true association signals, while adjusting for the required multiple testing.

Novelty of the variants: the observed associations were used to query the NHGRI-EBI GWAS catalog and the Phenopedia and Genopedia from HuGE Navigator (231). The phenopedia database was queried by disease (search terms: “systemic sclerosis”, “systemic lupus erythematosus”, “rheumatoid arthritis”, and “myositis”) and the genopedia by gene name based on the annotations from SNPnexus and previously described gene prioritization. Variants were classified as “new” if they had never been associated before with a single disease at the genome-wide significance level.

Functional enrichment analysis: to characterize the functional, cellular and regulatory contribution of the associated SNPs, we conducted a non-parametric enrichment analysis with GARFIELD (232). This software combines summary statistics with functional annotations from the ENCODE (233) and Roadmap Epigenomics (158) projects by calculating a fold enrichment, and assessing their significance via permutation testing while accounting for LD, MAF and local gene density. Functional annotation included genetic elements (GENCODE), DNase hypersensitivity sites (DHS), transcription factor binding sites (TFBS), histone modifications and chromatin states.

To determine whether any of the lead variants was an eQTL the online tools HaploReg v.4.1 (202) and the GTEx project (206) were queried. Additionally, the Capture HiC plotter (234) tool was used to display physical interactions between restriction fragments containing the variants and the gene promoters. This catalog of promoter-interacting regions aims to explain how the transcriptomic control of the genome works. The query was performed using the two promoter-capture datasets contained in the tool and their different cell types. Only the interactions between associated variants and reported genes affected by eQTLs were carried forward in the analyses. If there was no overlap with the genes modulated by the eQTLs, the resulting interacting genes were flagged as putative candidates for the diseases. The five new shared risk SNPs were queried as well and several putative candidates were found (**Supplementary Figure S3.1**).

Drug Target Enrichment Analysis: in order to assess if associated genes were enriched in drug targets, the target genes of eQTLs that overlapped with pCHi-C data were used to model a PPI network using STRING V.11.0 (203). When possible, the interaction confidence score was set in 0.9, which is the highest score calculated as a combined probability from different evidences of interactions corrected for the probability of observing an interaction by random chance. These protein products from the risk genes and those in direct PPI with them were then used to query the OpenTargets Platform (235) for drug targets. Additionally, the same platform was searched for drugs indicated, or in different phases of drug development, for the treatment of SSc, SLE, IIM and RA. Then, Fisher's exact test was used to calculate if the results of the meta-analysis were significantly enriched in pharmacologically active drug targets. To such purpose, we used gene-products related and unrelated to the analyzed disease, drug targets indicated for the disease and coding genes of the genome that are potentially druggable. The Drugbank database (236) and ClinicalTrials.gov (https://clinicaltrials.gov/) were searched for drugs that are currently in development for the treatment of RA and for information on publicly and privately supported clinical studies on SSc, SLE and IIM.

3.2. Results

3.2.1. Cross-disease meta-analysis and disease contribution

Following sample QC and imputation, a total of 11,678 cases and 19,704 non-overlapping controls were included in the genome-wide meta-analysis of 6,450,125 SNPs across the four diseases. The mean concordance rate among imputed and true genotypes was 0.999±0.0003. The final λ showed minimal evidence of population stratification in the meta-analysis (λ=1.025). Moreover, we calculated λ1,000 with consistent results (λ1,000=1.025). Summary of sample/variant QC are shown in **Supplementary Table S3.1**.

The global meta-analysis revealed 42 non-HLA significantly associated loci. Subsequent conditional analyses showed that 26 SNPs were independent (**Figure 3.1**). Seventeen variants were meta-analyzed under a fixed effects model, whereas 9 with random effects based on study heterogeneity.

To comprehensively explore the combinations of diseases contributing to the associations we applied a subset-based meta-analysis implemented in ASSET. Our model search yielded 26 SNPs associated with at least two IMIDs (**Table 3.1**). All of these variants were imputed in at least one dataset.

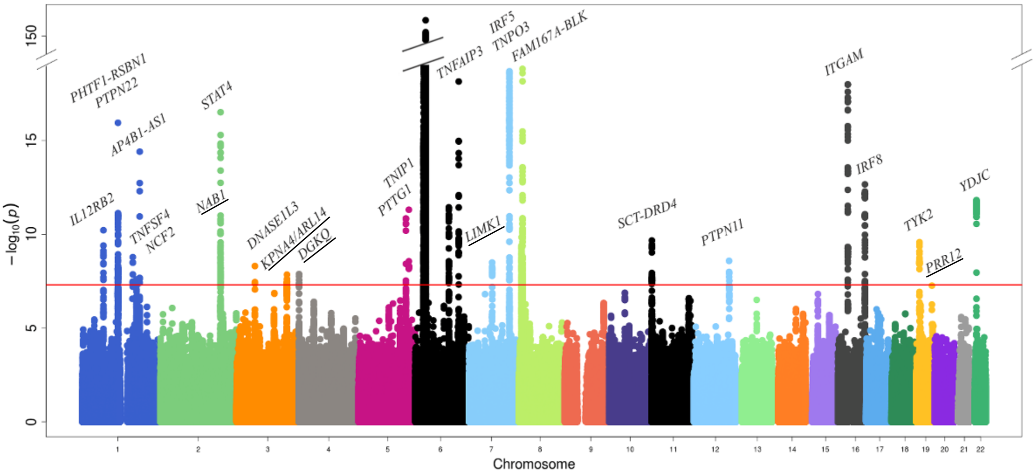


Figure 3.1. Meta-analysis results for the four systemic immune-mediated inflammatory diseases (IMIDs). The red line depicts the genome-wide significance threshold (*p*-value=5×10^{-8}). A total of 26 SNPs were independently associated with at least 2 systemic IMIDs. Most of the signals map to known susceptibility loci in autoimmunity (eg, *PTPN22, STAT4, TNPO3, FAM167A-BLK*) and five loci represent new shared susceptibility factors across IMIDs (underlined).

Table 3.1. Twenty-six independent variants associated at the genome-wide significance level (*p*-value $<5\times10^{-8}$) in the meta-analysis.

Chr	Position[a]	SNP	Gene[b]	Functionality[c]	Effect Allele	OR (CI 95%)	Meta-Analysis *p*-value[d]	Cochran's *p*-value	Contributing Disease[e]
1	67802371	rs6659932	*IL12RB2*	Intronic	C	0.85 (079-0.91)	6.08E-11	1.02E-02	**IIM, SLE**, SSc
1	114303808	rs6679677	*PHTF1-RSBN1*	Intergenic	A	1.34 (1.21-1.49)	2.30E-28	2.14E-04	**IIM, RA**, SLE
1	114377568	rs2476601	*PTPN22*	Coding (missense)	G	0.75 (0.67-0.83)	1.74E-28	1.06E-04	IIM, RA, SLE
1	114433946	rs1217393	*AP4B1*	Intronic	A	0.89 (0.85-0.92)	5.21E-09	4.91E-01	**IIM, RA**, SLE, **SSc**
1	173337747	rs2422345	*TNFSF4-LOC100506023*	Intronic	A	1.11 (1.05-1.18)	2.55E-08	6.00E-03	**IIM**, SLE, **SSc**
1	183532580	rs17849502	*NCF2*	Coding (missense)	T	1.36 (1.16-1.59)	3.93E-15	2.84E-04	**IIM**, SLE
2	191564757	rs744600	*NAB1**	3'Downstream	T	0.88 (0.85-0.92)	7.07E-11	7.60E-01	**IIM, RA, SLE, SSc**
2	191933283	rs13389408	*STAT4*	Intronic	C	1.27 (1.20-1.34)	3.10E-17	3.99E-01	**IIM**, SLE, SSc
2	191973034	rs10174238	*STAT4*	Intronic	A	0.73 (0.67-0.80)	2.76E-42	4.31E-07	**IIM**, SLE, SSc
3	58183636	rs35677470	*DNASE1L3*	Coding (missense)	A	1.22 (1.14-1.30)	4.96E-09	6.78E-01	**IIM**, SLE, SSc
3	160312921	rs112846137	*KPNA4-ARL14**	Intergenic	T	1.27 (1.17-1.37)	1.42E-08	9.55E-01	**IIM, RA, SLE, SSc**
4	965720	rs13101828	*DGKQ**	Intronic	G	1.11 (1.07-1.16)	1.32E-08	2.29E-01	**IIM, RA, SLE, SSc**
5	150438477	rs4958880	*TNIP1*	Intronic	A	1.16 (1.10-1.22)	1.45E-11	2.61E-01	**IIM, RA**, SLE, SSc
5	159887336	rs2431098	*PTTG1-MIR3142HG*	Intergenic	G	1.12 (1.05-1.20)	4.91E-12	1.42E-01	SLE, **SSc**

SNP single nucleotide polymorphism, *OR* odds ratio, *CI* confidence interval.

Table 3.1. Twenty-six independent variants associated at the genome-wide significance level (p-value $<5\times10^{-8}$) in the meta-analysis (continuation).

Chr	Position[a]	SNP	Gene[b]	Functionality[c]	Effect Allele	OR (CI 95%)	Meta-Analysis p-value[d]	Cochran's p-value	Contributing Disease[e]
6	106569270	rs802791	*PRDM1-ATG5*	Intergenic	C	0.87 (0.83-0.92)	3.65E-12	1.13E-01	SLE, SSc
6	138243739	rs58721818	*TNFAIP3*	3'Downstream	T	1.64 (1.46-1.84)	4.64E-23	1.65E-01	**IIM**, SLE, SSc
7	73537902	rs193107685	*LIMK1**	3'Downstream	C	1.52 (1.27-1.83)	3.21E-09	1.18E-01	**RA, SLE, SSc**
7	128589633	rs10954214	*IRF5*	3UTR	T	1.18 (1.13-1.23)	6.63E-17	3.64E-01	**IIM,** RA, SLE, SSc
7	128647942	rs13238352	*TNPO3*	Intronic	T	1.44 (1.30-1.60)	1.47E-38	2.12E-01	SLE, SSc
8	11341880	rs2736337	*FAM167A-BLK*	Intergenic	C	1.23 (1.17-1.30)	4.86E-22	1.29E-01	**IIM,** RA, SLE, SSc
11	633689	rs7929541	*SCT-DRD4*	Intergenic	G	0.89 (0.83-0.95)	2.14E-10	4.98E-04	**IIM, RA**, SLE, **SSc**
12	112871372	rs11066301	*PTPN11*	Intronic	T	1.11 (1.07-1.15)	4.20E-08	5.86E-01	**IIM, SLE, SSc**
16	85994484	rs35929052	*IRF8*	Intergenic	T	0.83 (0.78-0.88)	1.71E-09	4.69E-01	**IIM**, SLE, SSc
19	10462513	rs11085725	*TYK2*	Intronic	A	0.88 (0.83-0.92)	2.65E-10	1.86E-01	**IIM**, SLE, SSc
19	50121274	rs76246107	*PRR12**	Intronic	G	1.28 (1.14-1.43)	3.36E-08	1.50E-02	**IIM, SLE, SSc**
22	21985094	rs5754467	*YDJC*	5'Upstream	G	1.20 (1.13-1.27)	1.24E-13	8.59E-02	**IIM**, RA, SLE, **SSc**

[a] According to NCBI build GRCh37/hg19.

[b] Variant localization based on the nearest gene.

[c] Functionality obtained from SNPnexus.

[d] Results of meta-analysis either under a fixed (Cochran's Q test p-value≥0.05) or a random effect (Cochran's Q test p-value<0.05).

[e] Disease contributing to the association observed by the subset meta-analysis method with ASSET.

The diseases for which this locus has never been reported before at genome-wide significance level are shown in boldface.

*Denotes novel loci in the study.

All the variants in the table were imputed in at least one of the 18 case-control collections.

Among these 26 associations we found several key players in autoimmunity; interestingly ten of these associations (38%) have never been reported before for SSc, eight (31%) for SLE and RA, and 20 (77%) for IIM. Remarkably, five SNPs have not been reported previously for any of the diseases under study and thus constitute new shared risk loci in systemic seropositive rheumatic IMIDs (**Table 3.1**). These five new genomic associations include: the rs744600 SNP in the 3' region of the NGFI-A binding protein 1 (*NAB1*) gene (OR for the T allele=0.88, 95%CI=0.85-0.92, *p*-value=7.07×10^{-11}), and the intronic SNP rs13101828 mapping in the gene Diacylglycerol kinase theta (*DGKQ*) (OR for the G allele=1.11, 95%CI=1.07–1.16, *p*-value=1.32×10^{-8}). Of note, both genes have been previously associated with a chronic autoimmune liver disease (237,238). The intergenic SNP rs112846137, maps between the genes Karyopherin subunit alpha 4 (*KPNA4*) and the ADP ribosylation factor like GTPase 14 (*ARL14*) (OR for the T allele=1.29, 95%CI=1.07–1.56, *p*-value=1.42×10^{-8}). Interestingly, the gene *ARL14* showed a suggestive association in a pharmacogenomic GWAS of response to methotrexate in RA patients (239). In addition, we observe the associated SNP rs193107685 located in the 3' region of the LIM domain kinase 1 *(LIMK1*) gene (OR for the C allele=1.52, 95%CI=1.27–1.83, *p*-value=3.81×10^{-9}). The protein encoded by this gene regulates actin polymerization, a critical process in the activation of T cells (240). Finally, the SNP rs76246107 is located in an intron of the gene Proline rich 12 (*PRR12*) (OR for the G allele=1.28, 95%CI=1.14–1.43, *p*-value=3.36×10^{-8}), which was associated with fibrinogen concentration (241), and is an active regulator of the inflammatory response (242).

3.2.2. Associated loci and their functional enrichment on regulatory elements

To assess whether the associated variants lie in coding and non-coding regulatory and cell-type-specific elements of the genome, we performed an enrichment analysis with GARFIELD. The results obtained showed marked enrichment patterns mainly in blood cells and skin cells, with 247 significant enrichments (*p*-value ≤ 5×10^{-05}) (**Supplementary Figure S3.2**). **Table 3.2** summarizes the main enrichment results. Specifically, we found that the majority of associated variants were enriched in DHS hotspots in blood, as depicted in **Figure 3.2**. This functional category included a repertoire of cells from the immune system, such as B-lymphocytes (fold enrichment (FE)=11.68, empirical *p*-value (*p*emp) < 1×10^{-05}), T-lymphocytes (FE=10.42, *p*emp < 1×10^{-05}), including T helper (FE=7.81, *p*emp < 1×10^{-05}), T CD8^{+} (FE=7.61, *p*emp < 1×10^{-05}), and natural killer (FE=11.36, *p*emp < 1×10^{-05}) cells, and monocytes (FE=8.99, *p*emp < 1×10^{-05}). In line with this enrichment, disease-associated SNPs were enriched in enhancers (FE=14.99, *p*emp < 1×10^{-05}), within TSS (FE=14.87, *p*emp < 1×10^{-05}), and on TFBS (FE=12.20, *p*emp < 1×10^{-05}) in the B-lymphocyte cell line GM12878. Additionally, the highest enrichment was observed in the histone modifications H3K9ac (FE=14.02, *p*emp < 1×10^{-05}), and H3K27ac (FE=10.81, *p*emp < 1×10^{-05}) in the B-lymphocyte cell line, which are positively associated with gene activation. Although these modifications are increased in the promoters of active genes, the latter has been shown to be associated with active enhancers (243). Moreover, enrichment was observed in H3K4me1,2,3 sites, which usually surround TSS and are also positively correlated with gene expression (243).

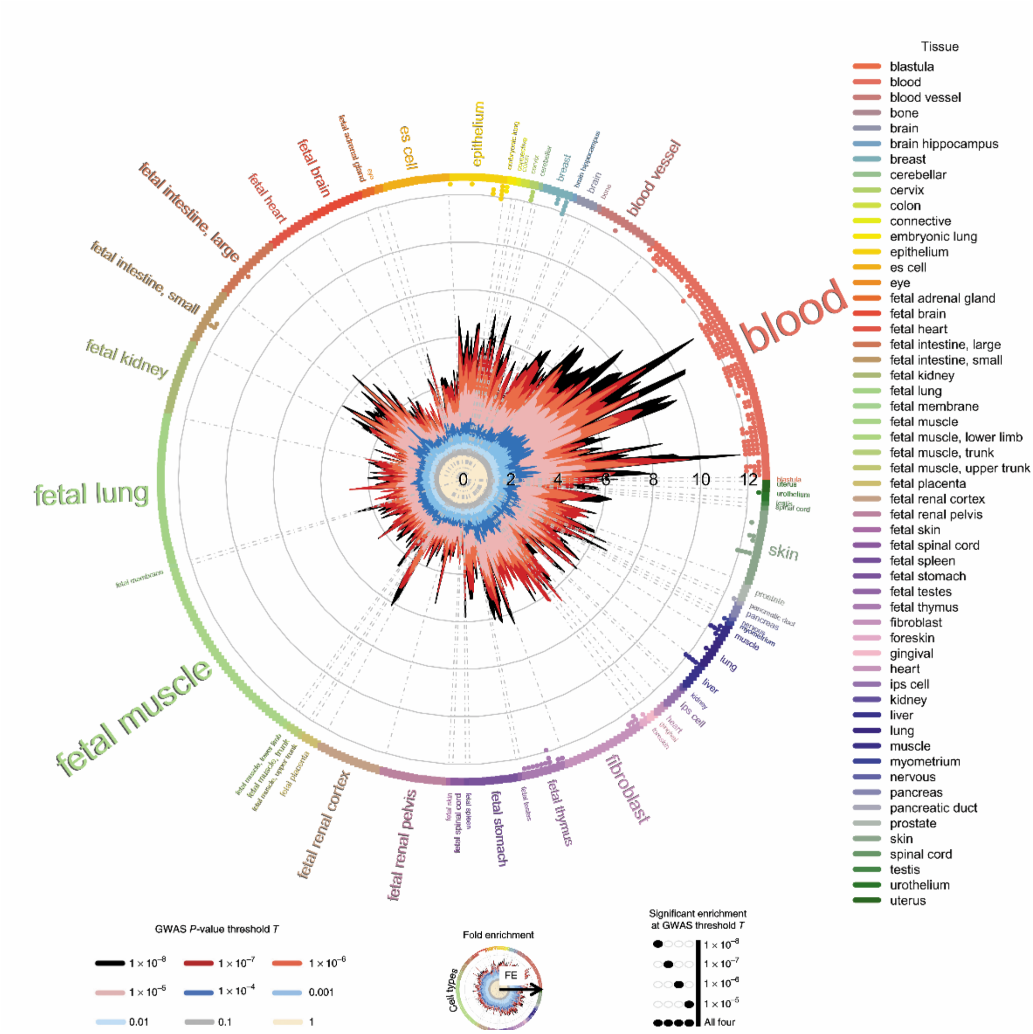


Figure 3.2. GARFIELD functional enrichment analyses in DNase hypersensitivity sites (DHS) hotspots. The wheel plot shows functional enrichment in systemic immune-mediated inflammatory diseases within DHS hotspot regions in encode and roadmap epigenomics. The radial axis depicts the FE calculated at different meta-analysis *p*-value thresholds. The font size is proportional to the number of cell types from the tissue, mainly enriched in blood cell types including a repertoire of immune cell lines.

Table 3.2. Summary of the most enriched functional annotations for the SNPs associated in the cross-disease meta-analysis at the genome-wide significance threshold (*p*-value <5x10^{-8}).

Category[a]	Tissue	Cell types	Type	NAnnot Thresh[b]	N Annot[c]	N Thresh[d]	Fold Enrichment[e]
Chromatin States	Blood	GM12878	Enhancer	13	10,944	33	14.99
		GM12878	TSS	12	10,182	33	14.87
Footprints	Blood	GM06990	Footprints	8	3,153	33	32.02
Histone modifications	Blood	GM12878	H3K9ac	21	18,903	33	14.02
		GM12878	H3K27ac	22	25,674	33	10.81
		GM12878	H2AFZ	22	25,824	33	10.75
		GM12878	H3K4me3	17	25,365	33	8.46
		GM12878	H3K4me2	23	34,807	33	8.34
		GM12878	H3K4me1	25	39,871	33	7.91
		GM12878	H3K79me2	16	25,683	33	7.86
Hotspots	Blood	GM06990	Hotspots	23	24,839	33	11.68
	Skin	NHEK	Hotspots	25	54,667	33	5.77
Peaks	Blood	GM06990	Peaks	13	6,433	33	25.50
TFBS	Blood	GM12878	TFBS	19	19,65	33	12.20

[a]Functional categories from the Encode and Roadmap Epigenomics.
[b]Number of LD-pruned annotated variants passing the meta-analysis threshold.
[c]Number of LD-pruned annotated variants in the reference dataset UK10K project.
[d]Number of LD-pruned variants passing the meta-analysis threshold.
[e]All enrichments with an associated empirical *p*-value < 1x10^{-5}
GM12878 B-Lymphocyte, *GM06990* B-lymphocyte, lymphoblastoid, *NHEK* Normal Human Epidermal Keratinocytes. *TFBS* transcription factor binding site, *TSS* transcription start site.

3.2.3. Expression quantitative trait loci (eQTL) and associated variants

In silico analysis of eQTLs revealed the role of 16 of the lead SNPs as eQTLs in whole blood, lymphoblastoid cell lines, transformed lymphocytes, skeletal muscle and transformed fibroblasts derived from European individuals from HaploReg v.4.1 (**Table 3.3**). Focusing on new associated variants, the T allele of the SNP rs744600 increases *NAB1* gene expression in lymphoblastoid cell lines (*p*-value=1.30x10^{-34}), and *HIBCH* expression in skeletal muscles (p-value=8.09x10^{-7}). The G allele of rs13101828 increases *DGKQ* expression in whole blood (p-value=3.29x10^{-45}), lymphocytes (p-value=5.23x10^{-19}), fibroblasts (p-value=4.44x10^{-6}), lung cells (p-value=8.42x10^{-28}) and several other tissues. The A allele of rs76246107 reduce *ALDH16A1* expression in lung cells (p-value=6.45x10^{-6}). Reassuringly, 14 of the 16 (87%) reported eQTLs showed a physical interaction between the SNP and the promoter of 15 of the genes affected by the eQTLs (**Table 3.3**), as suggested by CHi-C data. These independent evidences propose a mechanistic approach to understand the modulation of gene expression.

3.2.4. Drug target enrichment analysis

Genetic associations have the potential to improve the rates of success in the development of new therapies. We assessed if the protein-products from disease associated eQTLs and their direct PPI partners were enriched with pharmacologically active targets. We identified as eQTLs and PPIs 608 proteins for SSc, 630 for SLE, 632 for IIM, and 413 for RA, based on data on drugs at any stage of development collected from the Open Targets Platform (235). Using this information, for SSc, we found that 23 out of 73 (32%) proteins are targeted by drugs being studied for the disease (OR=16.80, p-value=1.41x10^{-18}). Similarly, 7 out of 25 (28%) proteins related to IIM and 13 out of 146 (9%) proteins related to SLE are addressed by drugs in consideration for IIM and SLE (OR=13.40, p-value=4.62x10^{-6}, OR=3.38, p-value=2.85x10^{-4}, respectively) (**Supplementary Table S3.2**).

On the other hand, we found that five of the loci identified in our meta-analysis interact with 17 genes that are considered drug targets, six of which are used for the treatment of these diseases (**Table 3.4**).

Table 3.3. Summary of the eQTL results in European samples for the SNPs independently associated in the meta-analysis.

SNP	Allele	Source	Gene	Tissue	*p*-value
rs6659932*	C	GTEx2015 v6	*IL12RB2*	Whole blood	3.72E-11
rs6679677*	A	Westra 2013	*PTPN22*	Whole blood	4.84E-10
rs2476601*	G	Westra2013	*PTPN22*	Whole blood	3.36E-10
rs1217393*	A	GTEx2015 v6	*AP4B1*	Skeletal muscle	5.45E-07
		GTEx2015 v6	*HIPK1*	Whole blood	7.71E-09
		Westra 2013	*PHTF1*	Whole blood	9.56E-05
		Westra 2013	*PTPN22*	Whole blood	2.67E-10
		Westra 2013	*RSBN1*	Whole blood	1.41E-10
rs744600*	T	GTEx2015 v6	*HIBCH*	Skeletal muscle	8.09E-07
		Lappalainen 2013	*NAB1*	Lymphoblastoid cell line	1.30E-34
rs13389408	C	GTEx2015 v6	*GLS*	Skeletal muscle	3.42E-09
		Westra 2013		Whole blood	2.98E-07
rs35677470*	A	GTEx2015 v6	*PXK*	Skeletal muscle	7.08E-06
rs13101828	G	GTEx2015 v6	*DGKQ*	Whole blood	9.28E-45
				Transformed lymphocytes	1.21E-25
				Transformed fibroblasts	9.78E-07
				Lung	8.42E-28
rs4958880*	A	Westra 2013	*TNIP1*	Whole blood	1.09E-03
rs10954214*	T	GTEx2015 v6	*IRF5*	Whole blood	2.56E-16
		Lappalainen 2013		Lymphoblastoid cell line	7.54E-31
rs13238352*	T	Lappalainen 2013	*IRF5*	Lymphoblastoid cell line	2.88E-13
rs2736337*	C	GTEx2015 v6	*FAM167A*	Whole blood	2.90E-26
			FAM167A	Transformed fibroblasts	1.90E-18
			FAM167A	Transformed lymphocytes	2.10E-15
			BLK	Whole blood	5.30E-13
rs2736337*	C	GTEx2015 v6	*BLK*	Transformed fibroblasts	1.30E-11
			BLK	Transformed lymphocytes	3.30E-06
rs7929541*	C	GTEx2015 v6	*TMEM80*	Transformed fibroblasts	1.22E-11
rs11085725*	T	GTEx2015 v6	*TYK2*	Whole blood	2.30E-06
			TMED1	Whole blood	8.80E-06
rs76246107*	A	GTEx2015 v6	*ALDH16A1*	Lung	6.45E-06
rs5754467*	G	GTEx2015 v6	*UBE2L3*	Whole blood	4.68E-06

New associated SNPs found in our meta-analysis are shown in boldface.
*Designates those SNPs where a physical interaction has been observed in promoter Capture Hi-C data in relevant immune cells.
SNP single nucleotide polymorphism.

Table 3.4. Summary of the plausible target gene products with drug indications in systemic IMIDs.

Associated SNP	Gene product	Association results[a]	Drugs[b]	Targets	Disease indication[c]
rs6659932	*IL12RB2*	**IIM, SLE**, SSc	Canakinumab	IL1B	RA
			Anakinra	IL1R1	RA
			Tofacitinib	JAK kinases	RA
rs13389408	*GLS*	**IIM**, SLE, SSc	Azathioprine	PPAT	RA, SLE
rs13101828	*DGKQ*	**IIM, RA, SLE, SSc**	Orlistat	LIPF	--
rs2736337	*FAM167A-BLK*	**IIM**, RA, SLE, SSc	Nintedanib	PDGFRB	SSc
			Dasatinib	BLK	--
			Imatinib	ABL1	--
			Osimertinib	EGFR	--
			Vandetanib	EPHA1	--
			Fingolimod	S1PR1	--
			Bosutinib	SRC	--
rs11085725	*TYK2*	**IIM**, SLE, SSc	Tofacitinib	JAK kinases	RA
			Tocilizumab	IL6R	RA, SSc
			Interferon Apha-2B	IFNAR1	--
			Idelalisib	PIK3CD	--
			Ruxolitinib	JAK1	--

[a] Based on our meta-analysis, diseases contributing to the observed association. The diseases where the association of this variant has never been reported before at genome-wide significance level are shown in boldface.

[b] Drugs from the OpenTarget platform with their corresponding target.

[c] Current indication of the reported drug. Non-immune mediated diseases were omitted.

SNP single nucleotide polymorphism.

3.3. Discussion

Genetic factors play an important role in the development of more than 80 IMIDs identified so far (244). Comorbidity of these diseases, increased familial clustering and shared risk variants have been widely documented (245). However, to date, these shared loci have been identified by simple comparison between studies, and just recently they have been determined by rigorous and systematic analysis (246). In the present study, we identified five unreported shared loci associated with systemic seropositive rheumatic IMIDs. This is the first large-scale meta-analysis, including more than 11,000 cases and 19,000 non-overlapping controls aiming to improve our knowledge regarding the genetic resemblances among these conditions. The present results will be discussed with a particular focus on their contribution to SSc pathogenesis knowledge.

Our results show that 85% of the associated variants were shared by at least three diseases. Interestingly, for several known RA susceptibility loci, the contribution of RA was limited. In this case, most of the associated variants were independent to the ones previously reported. Among the new associated SNPs, the signals mapping to *NAB1*, *DGKQ* and *KPNA4-ARL14* were associated to all of the diseases under study, whereas those mapping *LIMK1* and *PRR12* were associated with three of the diseases, excluding IIM and RA, respectively. NAB proteins are known to interact with early growth response family members and act as corepressors induced by type I IFN (247). In addition to SSc, the IFN signature has been previously implicated in these diseases (248–250). Interestingly, two IFN regulatory factors (*IRF5* and *IRF8*) previously associated to the conditions under study, were associated in the meta-analysis. Additionally, the associated SNP mapping to *NAB1* is an eQTL in lymphoblastoid cell line, which evidences its role in disease pathogenesis. Furthermore, the SNP associated in our study is correlated with the variant rs16832798 in European population, which has been recently associated to SSc in a large-scale meta-GWAS, pointing to *NAB1* as the most probable candidate gene (106). The DGKQ protein mediates cell signal transduction and can indirectly enhance the epidermal growth factor receptor signalling activity (251). This pathway regulates cell proliferation and migration, and its expression is augmented in the vasculature of patients with SSc with pulmonary involvement (252). Moreover, the risk allele was associated with an increased expression of the gene in lymphocytes, fibroblasts and lung. In the same line as *NAB1*, the association of *DGKQ* has been recently described in SSc by López-Isac *et al* (106) through the SNP rs11724804, which is in high LD with the index SNP in our study. In addition, this gene was associated with Sjögren's syndrome, a related connective tissue disease (253). Remarkably, rs193107685 and rs112846137 interact physically with the promoters of the *LIMK1* and *ARL14* genes, respectively, in dendritic cells (**Supplementary Figure S3.1**), which supports the increasingly important contribution of dendritic cells in the pathogenesis of SSc (254). The protein encoded by the gene *ARL14* is a GTPase involved in the recruitment of MHC class II containing vesicles and control the movement of dendritic cells along the actin cytoskeleton (255). In the same way, the protein encoded by *LIMK1* regulates many actin-dependent processes, including the assembly of the immune synapse between T cells and antigen presenting cells (256), an expected biological process involved in seropositive IMIDs, as well as cytoskeletal dynamics and cytotoxicity in NK cells (257). Finally, the gene *PRR12* has been previously associated with fibrinogen concentrations (242). Fibrinogen is considered a high-risk marker for vascular inflammatory diseases and is considered an accurate predictor of cardiovascular diseases (242,258). Moreover, this molecule is an active player in the coagulation cascade, responsible for the spontaneous formation of fibrin fibrils. In this regard, cardiovascular events and fibrosis are the most life-threatening complications described in SSc, IIM and SLE (5,259,260).

In addition to these new associated loci reported in all included diseases, there are also 6 other loci that have never been associated with SSc before the performance of this study. These loci correspond to *AP4B1, TNFSF4-LOC100506023, PTTG1-MIR3142HG, SCT-DRD4, PTPN11,* and *YDJC.* Among these, the association of the *TNFSF4-LOC100506023* locus have subsequently been confirmed by López-Isac *et al* (106). This locus has previously been suggested to be associated with SSc through different candidate gene studies (261,262), especially in subsets of patients positive for lcSSc and ACA, but it never reached the genome-wide level of significance. In this sense, 3 out of 11 (27.2%) loci associated for the first time with SSc in this study, have been confirmed in the largest meta-GWAS of SSc performed to date (106). Thus, it would be of great importance the replication of the other associated loci in order to confirm new susceptibility genes that could be implicated in SSc pathology. For example, *PTPN11*, which encodes the protein-tyrosine phosphatase non-receptor type 11, represents a good candidate gene as tyrosine phosphatases play a key role in the immune dysregulation of SSc (263). Concretely, *PTNP11* is involved in IL-4 signaling, one of the main secreted cytokines by activated Th2 cells in SSc, leading to activation of fibroblasts and their differentiation to myofibroblasts (264). It is also worth mentioning that the rs5754467 SNP on the *YDJC* locus represents an eQTL of *UBE2L3*, a susceptibility gene for multiple autoimmune diseases which encodes the ubiquitin-conjugated enzyme that facilitates activation of proinflammatory NF-κB signaling (265).

The associated SNPs observed in our study are highly enriched in functional categories in B and T cells, natural killer and monocytes, highlighting the relevance of these cells in systemic seropositive rheumatic

IMIDs. Beyond whole blood, the skin is the other tissue with significant functional categories, which is not surprising given the nature of these connective tissue diseases. Moreover, epithelial cells could transdifferentiate into mesenchymal cells and eventually contribute in fibrotic processes (15). In fact, patients with SSc are usually stratified according to the extent of skin involvement. On the other hand, the histone modifications observed are consistent with the ones reported in previous studies, where histone hyperacetylation have been described in synovial tissues in RA, in B cells in SSc and in CD4$^+$ T cells in SLE (243). Finally, the independent associated SNPs act as eQTLs in relevant tissues (**Table 3.3**) and *in silico* data from pCHi-C experiments showed the potential mechanisms in which most eQTLs modulate gene expression. Interestingly, all new associated SNPs interact with the promoters of surrounding genes, suggesting them as putative candidates with a role in the pathophysiology of these conditions.

The prevalence of SSc, SLE and IIM is low and there are no specific treatments for these diseases in comparison with RA. In the case of SSc, Tocilizumab is the only biologic drug approved for its use in SSc-associated interstitial lung disease (266). Therefore, given our current knowledge on the use of genetic findings in drug target validation and drug repurposing, we evaluated if drugs currently indicated for RA had the potential to be used in any of the other IMIDs under study. Our meta-analysis revealed that ten loci overlap with known RA risk genes. For instance, the gene-product of *TYK2* is targeted directly by Tofacitinib, which inhibits janus kinases (https://www.drugbank.ca/drugs/DB08895) or indirectly through the IL-6 family signalling pathway by targeting the IL-6 receptor with Tocilizumab (https://www.drugbank.ca/drugs/DB06273). Both drugs are currently indicated for patients with moderate to severe RA who respond poorly to disease-modifying antirheumatic drugs. As *TYK2* is associated with SSc, SLE and IIM, it is a good candidate for therapy repositioning in these diseases. As a proof of concept, Tofacitinib is currently on trial for SLE (clinical trial identifier: NCT03288324), SSc (NCT03274076) and dermatomyositis (NCT03002649). Overall, we found that five of the loci identified in our meta-analysis interact with 17 genes that are considered drug targets, six of which are used for the treatment of these diseases (**Table 3.4**). Another interesting candidate for drug repurposing is Imatinib, a kinase inhibitor that targets ABL1, which interacts with the gene product of *BLK*, a known locus associated with SSc and RA (**Table 3.4**). Imatinib is currently being tested for SSc (NCT00555581) and RA (NCT00154336).

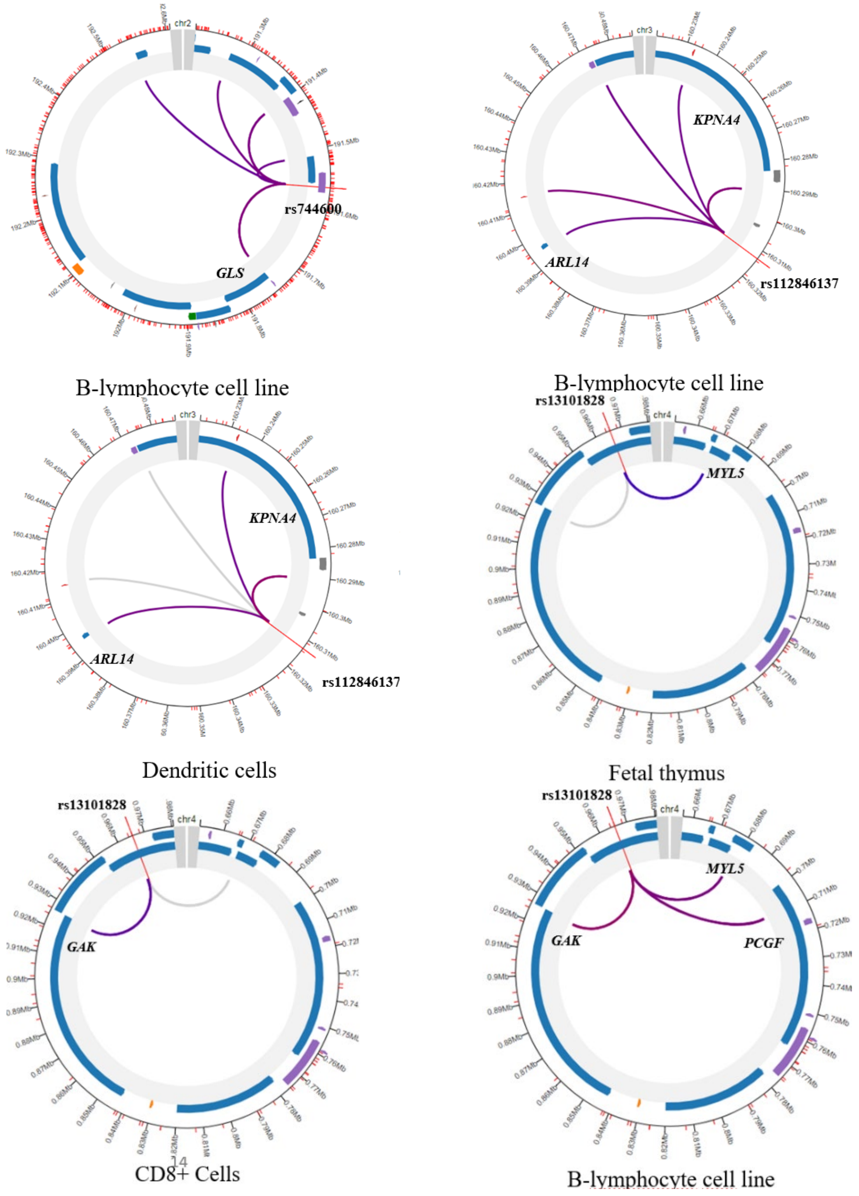
As compared with previous cross-disease studies of ADs, our study has the strength of analysing systemic seropositive rheumatic diseases, which is a more consistent clinical phenotype those previously investigated, where mixed seropositive and seronegative disease as well as systemic and organ-specific conditions were combined (114,115). The study of a more homogenous phenotype allowed us to determine that the type I IFN signalling pathway and its regulation play a more prominent role in these conditions than in others, based on the associations observed in *NAB1*, *TYK2*, *PTPN11*, *IRF5* and *IRF8*. Additionally, we performed a genome-wide scan to identify shared genetic aetiologies, as opposed to the study performed by Ellinghaus *et al* (114)*,* whose analyses were limited to the 186 autoimmune disease-associated loci implemented in the Immunochip platform. The study performed by Li *et al* (115), which was also a meta-analysis of GWAS data, was focused on paediatric autoimmune diseases, whereas our study was on a new combination of diseases in adult population. In addition, there is a posterior meta-analysis of Immunochip data performed in our laboratory incluing four seropositive diseases; including SSc, RA, celiac disease, and type 1 diabetes (113). In this study, authors identified 38 shared risk variants, many of them overlapping our results, including two of the new shared loci observed in our study (*NAB1* and *YDJC*), as well as many other shared loci associated with autoimmunity, such as *PTPN22, NAB1, STAT4, DNASE1L3, IRF5* or *TYK2*, to mention but a few. These results partially confirm the associations observed in our study.

In summary, this is the first study to investigate shared common genetic variation in four systemic seropositive rheumatic IMIDs in adults. We identified 26 genome-wide significant independent loci associated with at least two diseases, of which five loci had not been reported before. The shared risk variants and their likely target genes are functionally enriched in relevant immune cells and significantly enriched in drug targets, indicating that it may assist drug repositioning among genetically related diseases based on genomics data.

3.4. Supplementary Data

3.4.1. Supplementary Figures

Supplementary Figure S3.1. Circular view of the interactions from the new shared risk SNPs with genes nearby obtained from Promoter Capture Hi-C data in relevant immune cell types. Interactions are displayed as connecting lines depending on the confidence of the interaction. Grey lines are below threshold in the tissue. Only genes with maximum interaction score are reported.



B-lymphocyte cell line

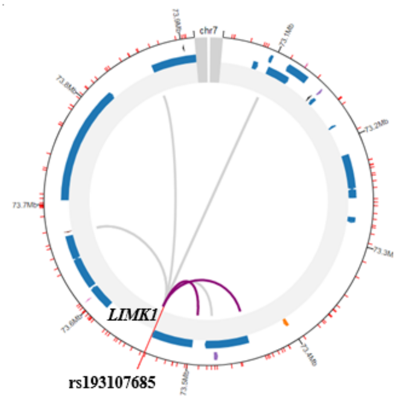
B-lymphocyte cell line

Dendritic cells

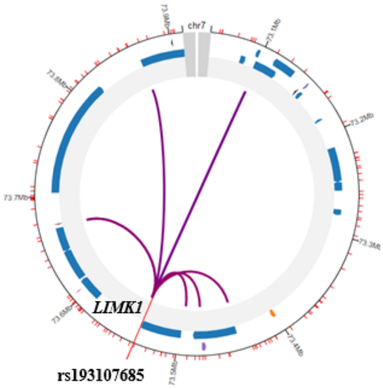
Fetal thymus

CD8+ Cells

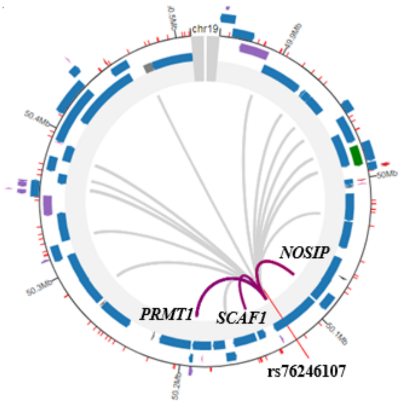
B-lymphocyte cell line



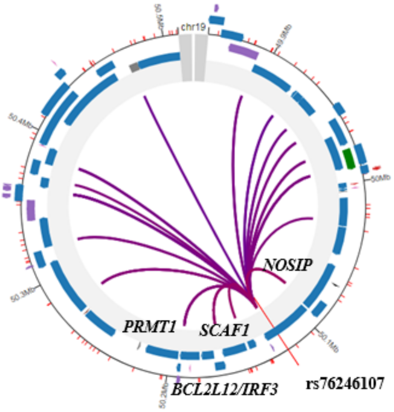
Dendritic cells



B-lymphocyte cell line

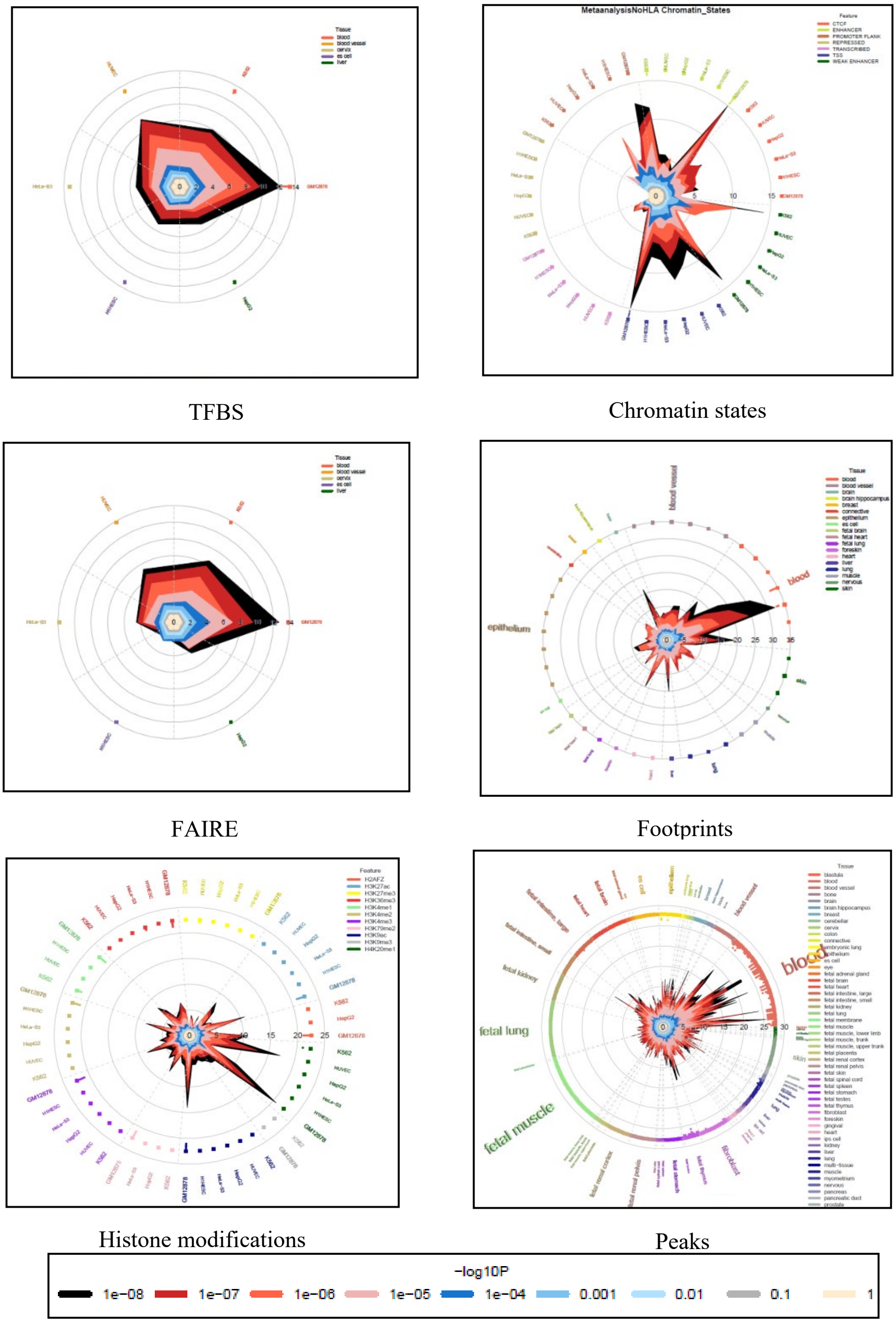


Dendritic cells



B-lymphocyte cell line

Supplementary Figure S3.2. Plots from the functional enrichment analysis with GARFIELD at different thresholds of *p*-values from the meta-analysis. Functional categories from the ENCODE project and Roadmap Epigenomics.



TFBS

Chromatin states

FAIRE

Footprints

Histone modifications

Peaks

3.4.2. Supplementary Tables

Supplementary Table S3.1. Summary of the study design, utilized samples, and analyzed variants in the study.

RA	N (Ca/Co) postQC	Genotyping platform	Genotyped SNPs after QC	Imputed SNPs	Utilized PC	Genomic inflation factor (λ)
Sweden	1893/744	Illumina 550K, Illumina 317K	293.631	7.618.025	2	1,02
United States	875/1201		446.637	7.589.337	5	1,03
United Kingdom	1827/1427		144.396	7.271.599	5	1,04
SLE						
Spain	458/499	Illumina HumanOmni1Quad BeadChip, Illumina HumanOmniExpressExome 8v1.2, HumanHap300v1.1, Illumina Human Hap550	727.382	7.370.411	2	1,03
Germany	199/1136		707.974	7.624.878	2	1,03
The Netherlands	270/1675		709.969	7.494.197	2	1,05
United Kingdom	985/1463		284.361	7.293.654	5	1,05
Italy	335/957		719.328	7.289.469	2	1,05
United States	907/3045		152.530	7.339.194	5	1,04
SSc						
Spain	362/362	Illumina Human CNV370K Beadchip, Illumina 550K	317.348	7.533.253	2	1,03
Germany	255/658		290.928	7.500.879	2	1,03
The Netherlands	182/631		286.266	7.528.731	2	1,02
United States	1482/2759		461.068	7.617.902	5	1,04
IIM						
Czech Republic/ Hungarian	316/242	Illumina HumanHap550 BeadChip, Illumina HumanCNV370-Duo v1 BeadChip, Illumina Human610-Quad v1 BeadChip, Illumina Human660W-Quad v1 BeadChip	226.643	7.369.026	2	1,01
Spain	47/253		226.034	7.518.585	2	0,84
United States	746/1152		220.533	7.551.455	5	1,05
Swedish/Dutch	120/326		228.315	7.560.114	2	1,00
United Kingdom	445/1177		227.556	7.470.059	2	1,01

N (Ca/Co): Number of cases and controls.

Utilized PC: Utilized principal components in the logistic regressions.

RA rheumatoid arthritis, *SLE* systemic lupus erythematosus, *SSc* systemic sclerosis, *IIM* idiopathic inflammatory myopathies, *SNP* single nucleotide polymorphism, *PC* principal component, *QC* quality control.

Supplementary Table S3.2. Drug target enrichment analysis.

	Systemic Sclerosis	**Systemic Lupus Erythematosus**	**Rheumatoid Arthritis**	**Idiopathic inflammatory myopathies**
Number of related gene-products[a]	608	630	413	632
Number of related gene-products & drug targets[b]	23	13	0	7
Number of unrelated gene-products & drug targets[c]	73	146	89	25
Number of unrelated gene-products & no drug targets[d]	20,012	20,012	20,012	20,012
Exact Fisher's test *p*-value	1.41E-18	2.85E-04	-	4.62E-06

[a] Target genes from the eQTL analysis and their protein-protein interaction network.

[b] Target genes from the eQTL analysis and their protein-protein interaction network that are drug target for any of the analyzed disease.

[c] Gene-products that are drug target for any of the analyzed disease but that are not related in our study samples.

[d] Gene-products that are not related in our study samples neither drug target for any of the analyzed disease.

Chapter 4: Expression quantitative trait locus (eQTL) analysis in systemic sclerosis identifies new candidate genes associated with multiple aspects of disease pathology

4.1. Material and methods

4.1.1. Study population

This study included 333 patients of European descent with a diagnosis of SSc according to the American College of Rheumatology/European league against rheumatism 2013 criteria (37) participating in the PRECISESADS project (https://clinicaltrials.gov/ct2/show/NCT02890134). A total of 524 controls without known AD were selected matching the cases according to age and sex. Patients and controls were recruited from 9 countries across Europe: Austria, Belgium, France, Germany, Hungary, Italy, Portugal, Spain and Switzerland. Patients and controls were randomly grouped into equal sized discovery and validation sets, matching for age, sex, and medication. **Table 4.1** describes the characteristics of both patient and control sets.

Table 4.1. Cohort charecteristics.

Variable	SSc		Controls	
	Discovery (n=167)	Validation (n=166)	Discovery (n=262)	Validation (n=262)
Females	85,60%	84,90%	78,20%	78,60%
Age, years	58.3±13.2	57.4±13.7	46.7±13.1	47.1±13.1
Age of Onset, years	48.5±13.6	48.9±14.2		
Disease duration, years	10.4±10.0	9.1±7.7		
Raynaud, n (%)	161 (96.4%)	160 (96.4%)		
Lung fibrosis, n (%)	51 (31.1%)	65 (39.1%)		
Intestinal symptoms, n (%)	58 (34.7%)	42 (25.3%)		
Calcinosis, n (%)	37 (22.1%)	33 (19.8%)		
PAH, n (%)	12 (7.2%)	11 (6.6%)		
Arthritis, n (%)	51 (30.7%)	45 (27.1%)		
Dyslipidemia, n (%)	51 (30.7%)	43 (25.9%)		
Sicca syndrome, n (%)	59 (35.3%)	43 (25.9%)		
Immunosuppressants, n (%)	39 (23.3%)	39 (23.5%)		
Prednisone > 5 mg/day, n (%)	39 (23.3%)	37 (22.3%)		
Chloroquien, n (%)	15 (9%)	16 (9.6%)		
Biologicals, n (%)	3 (1.8%)	3 (1.8%)		
No medication, n (%)	102 (61.1%)	103 (62.0%)		
Multiple medications, n (%)	28 (16.8%)	27 (16.3%)		

PAH pulmonary arterial hypertension, *SSc* systemic sclerosis.

All the patients and controls gave written informed consent for the study that was approved by local ethic committees according to standards set by International Conference on Harmonization and Good Clinical Practice (ICH-GCP) and to the ethical principles that have their origin in the Declaration of Helsinki (2013).

4.1.2. RNA sequencing

Total RNA was extracted from blood samples using Tempus Spin technology (Applied Biosystems) and depleted of alpha- and beta-globin mRNAs using globinCLEAR protocol (Ambion) and 1 µg of total RNA as input. 400 ng of globin-depleted total RNA was used for library synthesis with TruSeq Stranded mRNA HT kit (Illumina). Libraries were quantified using qPCR with PerfeCTa NGS kit (Quanta Biosciences), and equimolar amounts of samples from the same 96-well plate were pooled. Four pools were clustered on a high output flowcell (two lanes per pool) using HiSeq SR Cluster kit v4 and the cBot instrument (Illumina). Subsequently, 50 cycles of single-read sequencing were performed on a HiSeq2500 instrument using HiSeq SBS kit v4 (Illumina). The clustering and sequencing steps were repeated for three runs to generate enough reads per sample. Raw sequencing data were preprocessed using the software bcl2fastq, FastQC tools (267) and Cutadapt (268) to remove 3' end nucleotides below 20 Phred quality score and extraneous adapters, additionally reads below 25 nucleotides after trimming were discarded. Reads were then aligned to the UCSC Homo sapiens reference genome (Build hg19) using STAR v2.5.2b (269) and 2-pass mapping with default alignment parameters. Gene level expression estimates (Transcripts Per Million, TPM and read counts) were produced using RSEM v1.2.31 (270). A sample would pass the RNA single QC if: (i) the number of reads mapped to the genes was more than 7 million and (ii) the RNA integrity number (RIN) value was higher than 7.

4.1.3. Genotyping

Genomic data were obtained using Illumina SNP-chip platforms HumanCore-12-v1, InfiniumCoreExome-24v1-2 and InfiniumCoreExome-24v1-3. Only SNPs typed on all three platforms were used for imputation and analysis. Samples were subject to strict quality filtering using PLINK V1.9 (170) (SNPs with genotyping call rate <98% and deviating from HWE with a *p*-value <0.001 were removed) and analyzed for ancestry and identity using Frappe (271), PLINK and REAP (272). Samples were removed if they had less than 55% European ancestry, and Pi_hat scores greater 0.5. Imputation was done on the Michigan Imputation Server and imputed using Minimac4 (273) with Eagle v2.4 (274) for phasing. Imputed data was quality filtered using a QC score cut-off of 0.7 and each data subset (SSc and controls) was filtered for MAF < 0.05 and HWE *p*-value < 0.001.

4.1.4. eQTL detection

Data preparation: RNA-Seq and genetic data were checked to exclude mismatching samples by sex prediction and genotype mismatches using an in-house pipeline. To assess sex concordance, sex prediction was done for genetic data by PLINK v1.9 and for transcriptome data using specific sexual chromosome gene fractions. Regarding genotype concordance, genotypes were obtained from the RNA-Seq datasets using GATK best practices for RNA-Seq variant calling. In average, ~1,500 SNPs obtained from RNA-Seq variant calling procedure were also present in the Illumina Genotyping platforms. The concordance between the genotype and RNA-Seq variant calling was sought to be the highest across all potential pairs of comparisons in order to pass the quality criteria (in general, samples that fulfilled the criteria presented more than 1.8 times concordance than the next second pair).

Normalization: In short, bias from the genetic background were removed by calculating residuals using the first 3 PCs of the genetic data as predictors for the RNA-Seq data. Then, the inter-sample correlation of these residuals was calculated. A dataset-specific number of PCs of the inter-sample correlation was used to remove confounding influences in a linear manner and calculate gene expression residuals. The number of PCs used (SSc: 18 CTRL: 20) was optimized to yield the maximal number of eQTLs per dataset. PCs that showed a genetic influence were not used for cleaning.

Analysis and validation: The analysis was limited to 4,539 candidate SNPs that showed at least a suggestive level of association to SSc (*p*-value < $1x10^{-5}$) in a recent meta-GWAS published by our group (106). All SNPs with high LD (>= 0.8) were added to the candidate SNPs summing 13,253 SNPs in total for eQTL analysis.

The matrix eQTL R package (275) was used for the analysis, fitting a linear regression model that tests the influence of the number of risk alleles on gene expression residuals obtained by correcting for potential non-genetic confounders through a strategy described by Westra *et al* (276) using principal components. Analyses were focused on the identification of *cis*-eQTLs in a window of 1 million base pairs around the TSS of a gene. eQTLs were identified for each group separately to avoid interaction effects. A False Discovery Rate (FDR) lower than 0.05 was applied to declare significant genetic effects on gene expression. eQTLs were considered as validated if found in two sets, using a stringent cutoff (FDR < 0.05) in one set and a nominal *p*-value cutoff (*p*-value < 0.05) in the other.

To create a “cross-validated” set of eQTLs, the same strategy as above was followed, only that here the first set were eQTLs found using all SSc samples and the second set were eQTLs detected from all control samples. Again, a cross-validated eQTL had to be detected at high stringency in one set

(FDR < 0.05) and at with at least nominal stringency in the other set (nominal *p*-value < 0.05). In the first pass eQTLs and eGenes were detected for SNPs associated to SSc. In the second pass we detected eQTLs for all SNPs within 1 Mb of distance to an eGene detected in the first pass, including SNPs with no association to SSc.

SSc specific eQTLs: SSc eQTLs were called "SSc specific" if the eQTL was found validated using both SSc subsets and not found in any of the two control datasets, the joined control dataset or the cross-validated dataset at a nominal cutoff level of 0.1. Although none of the "SSc specific" eQTLs were found directly in public eQTL databases from healthy subjects (206,276), 27% of these eQTLs had proxy SNPs which were found with their respective gene in one of these databases and were therefore no longer considered SSc-specific.

4.1.5. Analysis of SSc patients without immunomodulating medication

For the subset of SSc patients which had no known medication, the analysis was repeated following the discovery and replication strategy described above (same selection of patients for the discovery and replication cohort and well-balanced subsets concerning sex, age, and medication) to find more SSc specific eQTLs. Beta values and *p*-values of both analyses (all patients vs patients without drugs) were highly correlated (r^2 > 0.95) as expected and only 15 eQTLs of 11,252 validated eQTLs (patients without drugs) were found new (with $FDR_{allSSCpatients}$>0.05). 11 of those 15 had $FDR_{allSSCpatients}$ < 0.065 and can be considered as missed by a small statistical margin, which leaves 4 eQTLs as new with $FDR_{allSSCpatients}$>0.23. Three of these eQTLs are associated to *DXO* expression and one to *TAMM41*. When analyzing SSc specific eQTLs using this subset of patients, 28 additional eQTLs were found for 10 genes, which were missed before by a very small margin, due to statistical cutoffs. Given the small changes and the lack of a "smoking gun pointing towards SSc" the whole analysis was not included in the results section as a separate point but rather included as a subpoint to the analysis of SSc specific eQTLs.

4.1.6. Stepwise linear regression and comparison of r-square values

Independent eQTL signals that influence the expression of a gene were determined following a stepwise linear regression procedure using 1) only eQTLs with an SSc associated SNP or 2) all eQTLs detected. Forward selection was repeated until no additional signal was detected at a nominal level of significance (p-value < 0.05). This was done for all SNP – eGene combinations where an eQTL was found in one or more of our datasets at a level FDR < 0.1. Stepwise linear regression was performed using a) only cases b) only controls or c) both. R-square values (called expression variance explained – EVE) were calculated for each gene for all of the combination above (1 & 2 & a/b/c) using the lm/summary function of R.

4.1.7. Differential expression analysis

The edgeR package (277) in R was used to calculate differential expression in 7 different cell types (neutrophils, monocytes, B lymphocytes, CD4$^+$ lymphocytes (memory and naive), CD8$^+$ lymphocytes and natural killer cells) using cellular composition of whole blood as a covariate as estimated from the expression profiles using CIBERSORT (278), using default parameters and no previous normalization. Additional covariates were disease, sex, age and medication and the interactions age:cells, medication:cells, disease:sex, disease:age. EdgeR was used with tagwise dispersion estimates and the function glmFit with robust=TRUE was used to calculate coefficients. Differential expression was determined by glmLRT. Differential expression for each cell type was visualized using the jtools v0.4.5 package in R.

Differential expression data from skin and lung tissues was obtained either from the publication tables (132,136) or by using the default analysis in GEO (279) :GEO2R with the dataset GSE58095 comparing all patients versus all controls.

4.1.8. Transcription factor binding site analysis

eQTLs in general are enriched for TFBS, and many TFBS show overlapping patterns. To restrict the number of eQTL SNPs for analysis, only SNPs which are part of the best expression models obtained by step-wise linear regression analysis explained above were analyzed. To allow for statistical fluctuations, SNPs which are in high LD (LD > 0.95) with the best expression model SNPs were included. The selection of SNPs to be tested for TFBS was called mSNPs. After obtaining all potential TFBS for these mSNPs we scored the effect of each mSNP on binding of the TF. As there is no gold standard for this score, TFBS enrichment for a range of scores was calculated. If the enrichment was significant at a level of FDR < 0.1 for at least three scores from the range, the overall enrichment of the particular TFBS was regarded as significant. To calculate enrichment, the Fisher test with a random selection of 50,000 eQTLs from the GTEx database V7 (206) was used as a background (matched for MAF and distance to TSS).

TFBSTools 1.24.0 (280) was used to detect TFBS with default settings using position weight matrices (PWM) from JASPAR2018 (281). A score which models how the different alleles of the mSNPs affects binding of the TF was calculated as positions within the TFBS are not of equal importance, as is reflected by the PWMs. Comparing the scores for the two alleles we can rank mSNPs for each TF. In detail, the information content (IC) was calculated for each position of the TFBS. The IC was then multiplied by the PWM to obtain a weighted score (ICWS) per base and position of the TFBS. Positions with high IC have usually a high weight for only one base in the PWM while positions with low IC have similar weights for each base. For each mSNP we calculated the ICWS for each allele and calculated their difference called DICWS. mSNPs with a high DICWS are more likely to affect the binding of their respective TF than mSNPs with a low DICWS.

There is no gold-standard cutoff for the DICWS and, therefore, we calculated the enrichment using a range of cutoffs of the DICWS. The cutoff was applied to both the mSNPs and to the random selection of 50,000 GTEx eQTL SNPs which was used as a background to calculate the Fisher test statistic. The random selection of GTEx eQTL SNPs, initially more than 50,000, was filtered to obtain a random selection which had the same MAF distribution as the mSNPs and the same distance to the nearest TSS distribution.

4.1.9. Drug-target analysis

2,384 different drugs and their 1,138 different target genes were extracted from the Open Targets database (235). Medication used for rheumatic and skin related diseases was extracted using the keywords "Rheumatic", "Arthritis", "Lupus", "Sjogren", "Scleroderma", "Dermatitis", "Psoriasis", "Arteriosclerosis", "Myositis", "Behcet", "Chondritis", "Spondylitis", "Gout", "Tendinopathy" from the same database leading to 542 drugs currently used to treat these diseases. Drugs are classified into six types (Antibody, Enzyme, Oligonucleotide, Oligosaccharide, Protein, Small molecule) and their mechanism of action enables to group their targets into receptors and kinases.

4.1.10. Tissue enrichment analysis

Some eQTLs have been found to be tissue and condition specific, while most eQTLs are found in many tissues under many conditions. An enrichment analysis will therefore detect significant enrichment/overlap between any two tissues. A baseline enrichment of blood eQTLs was calculated for all tissues separately using the GTEx database V7. The enrichment of the blood eQTLs obtained in this study was tested to observe if it was even higher as the baseline enrichment for all tissues using a z-test. In detail, the baseline enrichment was calculated using Fisher's test and the data from the GTEx V7 meta-analysis using a mvalue cutoff > 0.8 to determine eQTLs present for each tissue. A Z-test was used to compare the difference between two ORs. The z-score was calculated z = delta / SE(delta) which is normally distributed. Here delta = difference of log odds and SE(delta) = sqrt(SE_1^2 + SE_2^2). SE_1 and SE_2 were calculated as sqrt(1/TP + 1/FP + 1/TN + 1/FN) with TP=true positive, FP=false positive, TN=true negative and FN=false negative as taken from the contingency table.

4.1.11. Pathway enrichment analysis

As described above, by comparison of their expression variance explained, eGenes were grouped and those with a) high or intermediate impact of SSc genetics and b) low impact of SSc genetics, were analyzed for enriched pathways using Gene Ontology analysis from www.innatedb.com. Biological processes were considered enriched if their multiple-testing corrected *p*-value was < 0.05.

4.2. Results

4.2.1. Study design, gene and eQTL numbers, and comparison to external datasets

We aimed to explore the *cis*-genetic effects of SSc associated risk loci on expression in SSc and control datasets to detect potential disease-specific eQTLs and to model gene expression variation for gene prioritization. Prioritized genes were analysed for SSc hallmarks and drug repurposing and selected eQTLs were analysed for TFBS and tissue enrichment. **Figure 4.1** gives an overview of all analyses performed.

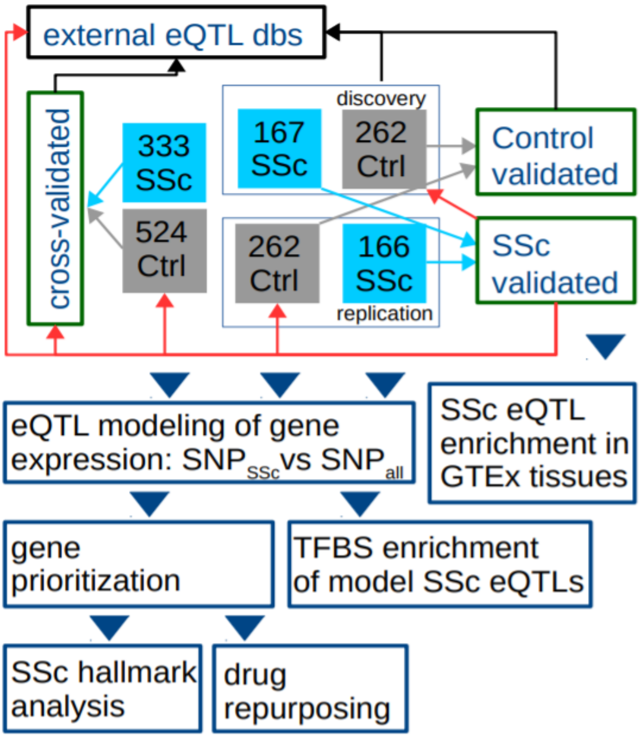


Figure 4.1. Experimental design. The upper part describes the different datasets obtained by our eQTL analyses (green boxes). The lower part (blue boxes) describes the flow of subsequent analyses performed with these datasets. Upper part: Healthy controls (Ctrl) and systemic sclerosis (SSc) subjects were equally split into discovery and replication set matched by sex, age and medication. eQTLs were called validated if found in two sets using a string end cutoff (FDR < 0.05) in one set and a nominal *p*-value cutoff (*p*-value< 0.05) in the other. eQTLs found using all SSc or all Ctrl subjects (at FDR < 0.05) were called "cross-validated" if found also in the other group (at nominal *p*-value < 0.05). Datasets were compared to external eQTL databases (depicted as black arrows) to determine the different levels of stringency of our setup. Comparisons made to detect potential SSc specific eQTLs are shown as red arrows. Lower part: eQTLs validated in SSc patients were tested for enrichment in GTEx tissues other than blood. All eQTLs obtained at nominal level (*p*-value<0.01) were used for stepwise-modeling (forward selection) to calculate the expression variance explained to prioritize genes, which were analysed for SSc hallmarks and potential drug repurposing. Model SSc eQTLs were used to calculate transcription factor binding site enrichment.

18,507 and 38,600 replicated *cis*-eQTLs were identified in SSc patients and controls, respectively, affecting the expression of 137 and 200 genes (eGenes), respectively. After validating across groups of eQTLs found in all SSc patients with eQTLs found in controls, and vice versa, a total of 49,123 eQTLs were identified, influencing 236 eGenes with a median of 73 eQTLs per gene. The maximal number of eGenes detected in any of the datasets at nominal level (*p*-value < 0.01) was 565, among them 64 long non-coding RNAs like *XXbac-BPG181B23*.7, *TAPSAR1* or *HCG11*.

The eQTLs of a) both discovery sets, b) cross-validated eQTLs, and c) the intersect of validated control and validated SSc eQTLs were compared against the GTEx database. We found 66%, 15%, and 8% unknown eQTLs, respectively, which depicts the different levels of stringency of our setup. Of interest, 95% of eQTLs in our whole blood dataset, which overlap GTEx, were found in multiple tissues according to GTEx.

4.2.2. SSc specific eQTLs

eQTLs replicated in SSc whole blood were compared to eQTLs observed in control datasets with low stringency (nominal *p*-value < 0.1). We found 59 eQTLs from 16 genes potentially specific to SSc. Repeating our analysis with a subset of patients, which did not receive immuno-modulating drugs, revealed 28 additional eQTLs and 6 additional genes. In depth comparison to known blood eQTLs from healthy controls (GTEx V7) (276,282) and their proxies (r^2>0.8) excluded 24 eQTLs (27%) from being SSc-specific. Careful examination suggested eQTLs from *HLA-B*, *NCR3*, *RAF1*, *NEU1*, *HLA-DQA1*, *HLA-DOB*, *HID1* and *IER3* to be the best candidates for SSc-specific eQTLs (**Figure 4.2**).

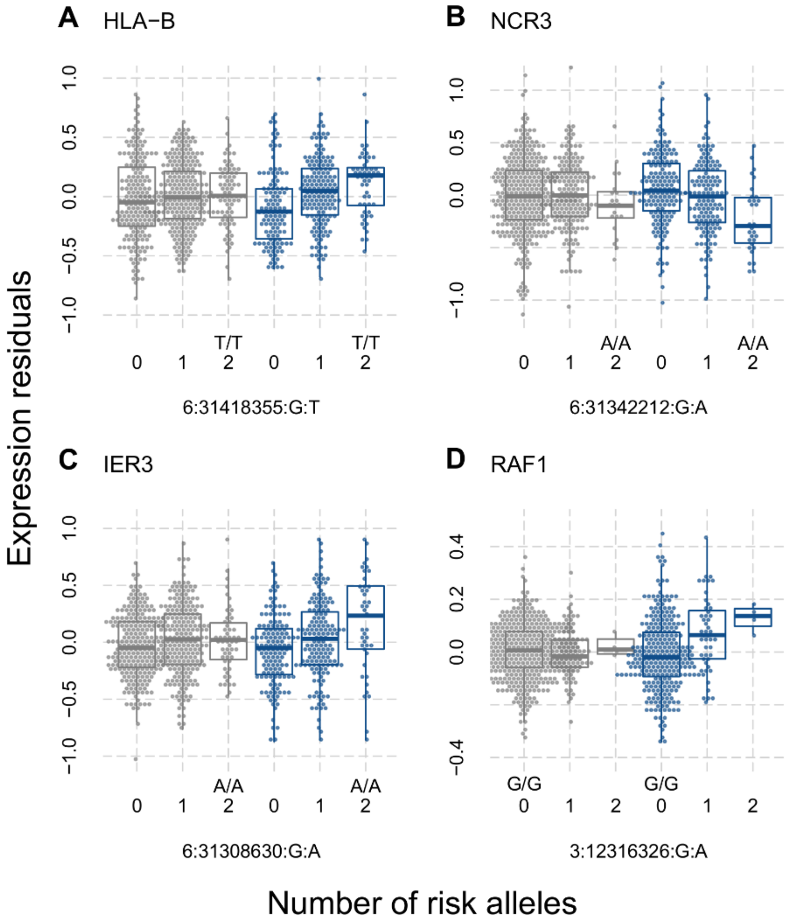


Figure 4.2. Expression quantitative trait loci found in patients with systemic sclerosis (blue) but not in controls (gray). Residual expression levels, determined using principal components analysis, of the genes *HLA– B* (A), *NCR3* (B), *IER3* (C), and *RAF1* (D) are shown for the indicated genotypes in controls and SSc patients. The number of minor alleles, the risk genotype, and single nucleotide polymorphisms are indicated on the x- axis. Data are shown as box plots. Each box represents the 25th to 75th percentiles. Lines inside the boxes represent the median. Lines outside the boxes represent the 10th and 90th percentiles. Dots represent individual subjects.

4.2.3. Enrichment of blood eQTLs in tissues affected by disease

We explored if the validated blood eQTLs can be interpreted in other contexts beyond immunity. The GTEx database provides a comprehensive overview of eQTL sharing among 49 different tissues. Using a meta-analysis published by GTEx V7, we found that only 6% of eQTLs are tissue-specific, 81% have been detected in at least 5 tissues, and 15% are present in more than 90% of tissues. This clearly shows that eQTLs detected in blood can be interpreted functionally in other tissues. Indeed, 95% of the GTEx-known eQTLs detected in this study are found in at least 10 different tissues apart from blood. We investigated whether the eQTLs identified in our study were enriched in the GTEx eQTLs of non-blood tissues to test our assumptions on interpretability beyond the context of whole blood.

A significant enrichment was found in 19 tissues (**Figure 4.3**), the majority of which can readily be interpreted in the context of SSc, as the disease affects many tissues like lungs, heart, and esophagus.

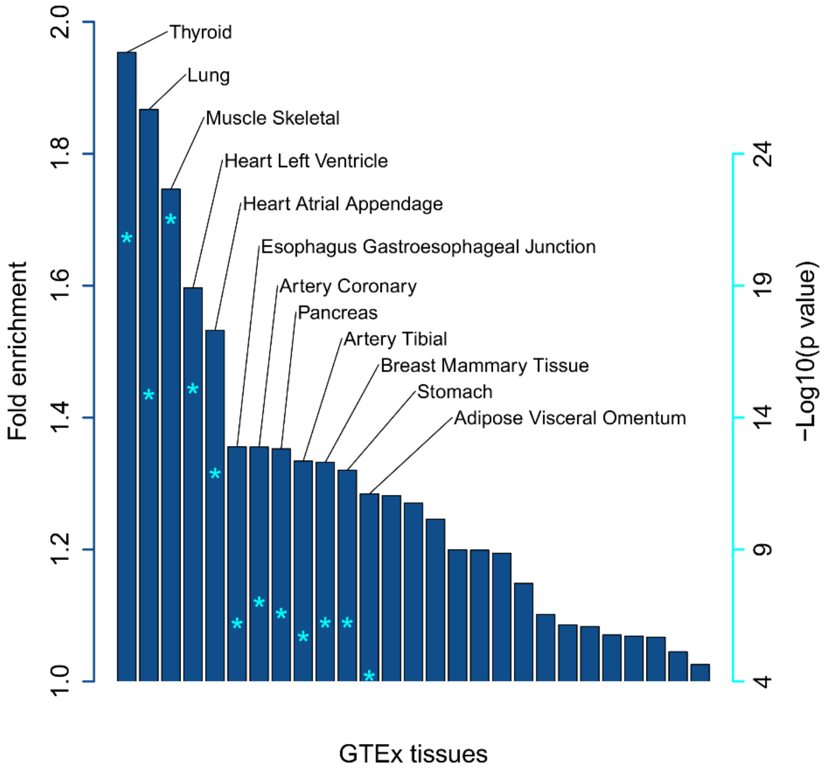


Figure 4.3. Enrichment of blood expression quantitative trait loci in disease-relevant tissues in patients with systemic sclerosis. Asterisks inside the bars indicate the level of significance adjusted for multiple testing (false discovery rate), corresponding to the values shown on the right.

4.2.4. Expression variance explained (EVE) can be used to prioritize SSc eQTLs and SSc eGenes

While many eGenes with an SSc-specific eQTL can probably explain the pathogenesis of SSc at least partially, we decided to focus on the candidate eGenes that are most affected by SSc genetics.

To measure the influence of genetics on gene expression, we used a stepwise modeling procedure to obtain independent eQTLs per gene and calculate the EVE. Comparing the EVE using only SSc-specific eQTLs ($\mathrm{EVE}_{\mathrm{SSc}}$) against the EVE using all eQTLs ($\mathrm{EVE}_{\mathrm{all}}$; including eQTLs unrelated to SSc) we obtain a measure (ratio) of how much EVE can be attributed to SSc genetics. As shown in **Figure 4.4A**, for 104 eGenes (18%) highlighted in red, the EVE differed by less than 30%. In orange, 130 of eGenes (23%) showed stronger differences in EVE, but with an $\mathrm{EVE}_{\mathrm{SSc}}$ still above 0.05. The remaining 331 eGenes depicted in blue had low $\mathrm{EVE}_{\mathrm{SSc}}$ (< 0.05), and the EVE differed by more than 30%. This comparison distinguished three groups with high, intermediate, and low influence of SSc genetics (**Figure 4.4A**).

Three groups of eGenes were identified based on the impact that SSc genetics had on their expression. We analyzed these groups for enriched pathways (FDR < 0.05), and biological processes from gene ontology, and found that 52% (122 of 233) of eGenes in the high- or intermediate-impact group were located in immune-related pathways, as compared to eGenes in the low-impact group (only 17% of eGenes). An in-depth review of the literature and gene ontologies helped us assign 66 and 31 eGenes to SSc-related biological processes linked to fibrosis and vasculopathy, respectively. Many of these eGenes belong to the high- or intermediate-impact group (**Figure 4.4B - D**). The eGenes on whose expression SSc genetics have an intermediate or high impact are most likely to shed light on the complex pathology of this disease.

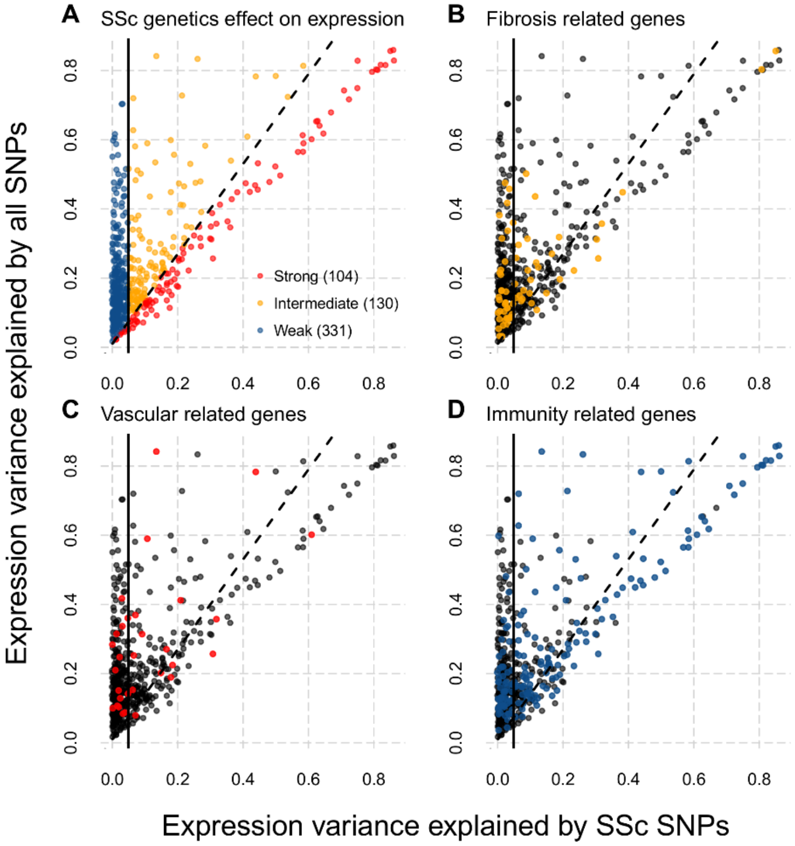


Figure 4.4. Gene expression variance explained by expression quantitative trait loci (eQTLs) can distinguish levels of influence of systemic sclerosis (SSc) genetics on expression and prioritize genes affected by eQTLs. The expression variance explained (r^2) by eQTLs associated with SSc in a recent genome- wide association study (using single nucleotide polymorphisms (SNPs) with association *p*-value < 10^{-5}) (106) was plotted against the expression variance explained by all eQTLs found within 1 Mb of a gene, whether or not they were associated with SSc. **A**, Groups of eGenes showing strong (red), intermediate (yellow), or weak (blue) influence of SSc genetics. **B– D**, Same eGenes as shown in **A**. Highlighted are eGenes related to **B**, fibrosis (yellow), **C**, vascular processes (red), and **D**, immunity (blue). The eGenes not related to any of these hallmarks are depicted in black.

4.2.5. SSc eGenes grouped by the hallmarks of SSc pathogenesis

Three features of SSc pathogenesis can be attributed to 134 of the 233 eGenes (58%) for which SSc genetics had an intermediate-to-high impact on expression, namely: alteration of immune response, fibrosis, and vasculopathy (**Table 4.2**). The genes implicated in innate and adaptive immune-cell processes represent the largest subgroup, with 122 eGenes. Interestingly, 27 HLA eGenes and 8 eGenes related to IFN pathways were identified, including important SSc-associated susceptibility loci dysregulated in SSc (142,205,283). Furthermore, there were 27 SSc eGenes associated with biological processes related to fibrosis, and 16 eGenes related to vasculopathy or angiogenesis. These pathways are considered to be potential targets of future disease-modifying therapies for SSc (284). Of interest, we also found 25 eGenes related to apoptotic processes, which support the hypothesis of a relevant role of apoptosis in SSc (285).

4.2.6. Differential expression of SSc eGenes in disease-affected tissues

Given that the SSc-specific eQTLs detected in whole blood were observed to be enriched in other tissues affected by the disease, we decided to analyze the expression of the prioritized 233 SSc eGenes in the skin, lungs, and seven blood cell types using public datasets (132,136) (GSE58095) and our whole blood dataset, with deconvolution of blood cell compositions. The data are presented in **Table 4.2**.

One hundred five SSc eGenes (45%) were found to be differentially regulated in one of the tissues investigated. A total of 57 SSc eGenes (24%) were downregulated in one of the three tissues investigated, whereas 55 SSc eGenes (24%) were upregulated. In addition, 40 SSc eGenes (17%) were differentially expressed in the skin of SSc patients. A total of 11 eGenes (5%) were found to be differentially regulated in the lung samples and lung fibroblast cultures of SSc patients. Differential expression analysis of seven blood cell types in SSc revealed 72 SSc eGenes (31%), most of which (99%) showed a consistent direction of regulation (up / down) in at least 5 cell types.

Table 4.2. Differentially expressed eGenes associated with hallmarks of SSc.

Gene	Impact of SSc genetics on expression	SSc Hallmarks			Differential expression ($\log_2$FC)*		
		Immunity	Fibrosis	Vascular	Blood	Skin	Lungs
AGER	high	+	-	+	-5.31	-	-
BLK	high	+	-	-	-	0.1	-
C2	high	+	-	-	-	0.45	-
C4A	high	+	-	-	-19.33	-	-
C4B	high	+	-	-	-19.57	-	-
CCHCR1	high	+	-	-	-4.58	-	-
CFB	high	+	-	-	-	0.4	-
DDAH2	high	+	-	+	-4.28	-	-
HLA-B	high	+	-	-	-4.49	-	-
HLA-DPA1	high	+	-	-	-	0.34	1.07
HLA-DQA1	high	+	-	-	-	-	1.04
HLA-DQB1	high	+	+	-	-	0.48	-
HLA-DRA	high	+	-	-	-	0.29	1.09
HLA-DRB5	high	+	-	-	-	-	1.25
HLA-DRB6	high	+	-	-	-	0.29	-
HSPA1B	high	+	-	-	-7.14	-	-
LST1	high	+	-	-	-5.72	0.23	-
LTB	high	+	+	-	-7.68	0.64	-
LY6G5C	high	+	-	-	-9.78	0.11	-
MICA	high	+	-	-	-6.29	-	-
MICB	high	+	-	-	-	0.21	-
NCR3	high	+	-	-	-9.71	-	-
NEU1	high	+	-	-	-	0.15	-
NOTCH4	high	+	+	+	-	0.23	-
RAB2A	high	+	-	-	-	-0.21	-
RNF5	high	+	-	-	-4.84	-	-
TAP1	high	+	-	-	-	-	1.23
TNXB	high	+	+	-	-7.01	-	-
AIF1	intermediate	+	-	-	-5.32	-	-
CCDC104	intermediate	+	-	-	-3.71	-	-
CD151	intermediate	+	-	-	-6.94	0.3	-
CD247	intermediate	+	-	-	-4.27	-	-

*Adjusted *p*-value < 0.1 for all values shown.
SSc systemic sclerosis, *FC* fold change.

Table 4.2. Differentially expressed eGenes associated with hallmarks of SSc (continuation).

	Impact of SSc genetics on	SSc Hallmarks			Differential expression ($\log_2$FC)*		
Gene	expression	Immunity	Fibrosis	Vascular	Blood	Skin	Lungs
CD40	intermediate	+	+	+	-	0.19	-
CTSB	intermediate	+	+	-	-	0.4	1.14
ELMO1	intermediate	+	-	-	5.56	-	-
ERAP1	intermediate	+	-	+	5.11	-	-
FLNB	intermediate	+	+	-	3.43	0.13	-
GTF2H4	intermediate	+	-	-	-	0.19	-
HLA-A	intermediate	+	-	-	-	0.25	1.06
HLA-DMA	intermediate	+	-	-	-	0.36	1.05
HLA-DMB	intermediate	+	-	-	-	0.32	1.05
HLA-DOA	intermediate	+	-	-	5.42	0.2	-
HLA-F	intermediate	+	-	-	-4.63	-	-
HLA-H	intermediate	+	-	-	-	0.23	0.99
HSPA1L	intermediate	+	-	-	-4.42	-0.14	-
IDUA	intermediate	+	+	-	-	0.26	-
IER3	intermediate	+	-	+	-	-	1.15
IFI30	intermediate	+	+	-	-3.78	-	-
MPI	intermediate	+	-	-	-2.73	-	-
MSRA	intermediate	+	-	-	-	0.15	-
PSMB8	intermediate	+	+	+	-4.49	-	-
PSMB9	intermediate	+	-	-		0.29	-
PXK	intermediate	+	-	-	2.91	-	-
RXRB	intermediate	+	-	-	-	0.15	-
SUMO2	intermediate	+	-	-	-	-0.21	-
TAPBP	intermediate	+	-	-	-	0.24	-
TNPO3	intermediate	+	-	-	5.64	-	-
TUBB	intermediate	+	-	-	-	0.16	-
UBE2L3	intermediate	+	-	-	-2.25	-	-
UNC119B	intermediate	+	+	-	2.33	-	-
CLIC1	intermediate	-	+	-	-2.9	-	-
FLOT1	intermediate	-	+	-	-4.9	0.28	-
PHF1	intermediate	-	+	-	-3.38	-	-
RPS18	intermediate	-	+	-	-9.34	-	-
SYNGAP1	intermediate	-	+	-	3.5	-	-
UQCC2	intermediate	-	+	-	-5.03	-	-

*Adjusted *p*-value < 0.1 for all values shown.
SSc systemic sclerosis, *FC* fold change.

4.2.7. Results of transcription factor binding site analysis

We investigated TFBS enrichment in SSc eQTLs. Only the independent eQTLs included in the models that best predicted eGene expression, as determined by stepwise linear regression, were included. Then, TFBS enrichment was estimated, as compared to genome-wide eQTLs from the GTEx database, to control for the fact that all TFBS motifs are highly enriched in eQTL sites in general.

Of the 537 TFBS profiles assessed (281), 24 (5%) were stably enriched in best-model SSc eQTLs. The TFs were of different classes, with five homeodomain TFs, four TFs of the T-box type, four C2H2 TFs, and two GATA TFs, to name but those with multiple members of the same class. Of the 24 TFs, we found ten and 16 TFs expressed in whole blood and skin, respectively, of which 5 TFs were differentially regulated (FDR < 0.1) in skin, lungs or blood cells from SSc patients (**Table 4.3**). *KLF4* and *ID4* were downregulated in skin, *TBX4* was upregulated in lungs, and *ELF1,* and *MGA* were upregulated in almost all the seven blood cell types assessed (**Figure 4.5**).

Table 4.3. Differentially expressed transcription factors with enriched binding sites in SSc- associated eQTLs in expression models.

		Differential expression (log_2FC)*		
Gene	**TF class**	**Blood**	**Skin**	**Lung**
ELF1	Ets	4.68	-	-
MGA	T-box	4.3	-	-
KLF4	C2H2 ZF	-	-0.36	-
ID4	bHLH	n.e.	-0.23	-
TBX4	T-box	n.e.	-	0.74

*Adjusted *p*-value < 0.1 for all values shown.
TF transcription factor, *FC* fold change, n.*e.* not expressed.

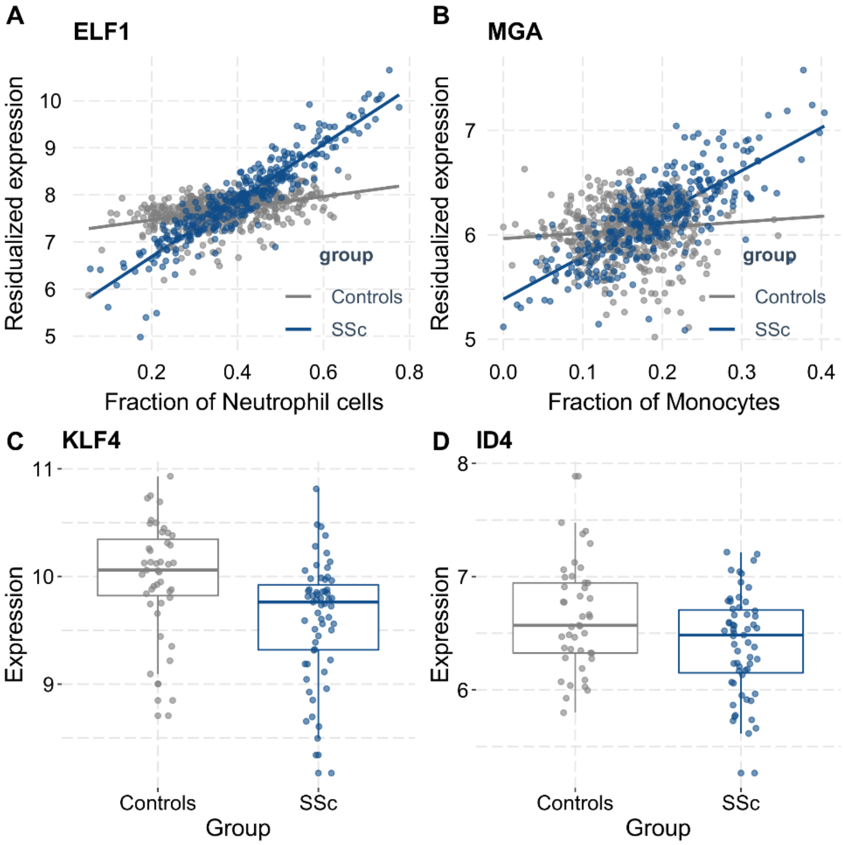


Figure 4.5. Differential expression of the transcription factors *ELF1*, *MGA*, *KLF4*, and *ID4* in patients with systemic sclerosis (SSc) compared to controls. **A** and **B**, Residual expression of *ELF1* in neutrophils (**A**) and *MGA* in monocytes (**B**) from controls and SSc patients. Values on the x- axis are the percentage of cells investigated per patient as obtained from the Cell- type Identification by Estimating Relative Subsets of Known RNA Transcripts (CIBERSORT) algorithm. *ELF1* and *MGA* were up- regulated in SSc patient tissues. **C** and **D**, Log$_2$ expression of *KLF4* (**C**) and *ID4* (**D**) in skin from controls and SSc patients. *KLF4* and *ID4* were down- regulated in SSc patient tissues. Data are shown as box plots. Each box represents the 25th to 75th percentiles. Lines inside the boxes represent the median. Lines outside the boxes represent the 10th and 90th percentiles. Dots represent individual subjects.

4.2.8. Drug repurposing

We explored whether any of the 233 eGenes prioritized in the present study encode target proteins of drugs being tested in ongoing clinical trials, as reported on the OpenTargets platform (235). We observed that 15 of the 233 eGenes (6.4%) overlapped with pharmacological targets of which *TNF* (NCT01670565), *BLK* (NCT00764309) and *TUBB* (NCT03198689) have been tested in clinical trials in SSc patients.

Next, we tested whether medications used for other immune-mediated diseases (105 antibody-targeted, 48 kinase inhibitor-targeted, and 195 receptor-targeted drugs; see material and methods) addressed the proteins coded by the SSc eGenes. We found five additional SSc eGenes: *LTA*, *LTB*, *IL12A*, *CD40* and *RXRB.* Further investigation beyond immune system-related targets revealed *ERAP1* and *ERAP2,* which can be addressed by aminopeptidase inhibitors.

Expression analysis in whole blood, skin and lung tissues revealed that six of the ten drug-target SSc-specific eGenes are differentially regulated in blood cells and/or skin of SSc patients (**Table 4.4**). *ERAP1* was upregulated in blood cells of SSc patients, whereas *LTB* was downregulated. *LTB*, *CD40*, *RXRB*, *BLK,* and *TUBB* were upregulated in the skin of SSc patients.

In summary, seven genes that have been considered for the treatment of conditions similar to SSc are potential candidates for study in clinical trials on SSc.

Table 4.4. Drug target genes with eQTLs detected for SNPs with association to SSc and their differential expression in tissues affected by SSc. These eGenes are candidates for clinical trials and drug repurposing.

Gene	**Likely impact of SSc genetics**	**SSc Hallmark**			**Differential expression (log_2FC)***		
		Fibrosis	**Vascular**	**Immunity**	**Blood**	**Skin**	**Lung**
BLK	high	-	-	+	-	0.1	-
CD40	intermediate	+	+	+	-	0.19	-
ERAP1	intermediate	-	+	+	5.11	-	-
ERAP2	intermediate	-	+	+	-	-	-
IL12A	high	-	-	+	-	-	-
LTA	high	+	-	+	-	-	-
LTB	high	+	-	+	-7.68	0.64	-
RXRB	intermediate	-	-	+	-	0.15	-
TNF	high	+	-	+	-	-	-
TUBB	intermediate	-	-	+	-	0.16	-

*Adjusted *p*-value < 0.1 for all values shown.
SSc systemic sclerosis, *FC* fold change.

4.3. Discussion

In this study, the integrated analysis of expression and genetic data in a large SSc cohort identified novel eQTLs in whole blood of SSc patients, which are enriched in disease-relevant tissues. We found 64 eQTLs potentially specific to SSc, which were not found in either our cohort of healthy controls or any of the public blood eQTL databases (GTEx V7) (276,282). This finding suggests that additional mechanisms exist that render these eQTLs active in disease and neutral in healthy subjects. The most likely explanation is the differential expression of transcription factors associated with a disease, as has been suggested previously (217,286). Indeed, we showed that of the 24 transcription factors associated with SSc by our analysis of TFBS enrichment, ≥5 were differentially expressed in disease-relevant tissues. The eQTL analysis of the most likely associated SSc risk loci, prioritizing genes (eGenes) where SSc eQTLs explain >5% of expression variance, led to a strong enrichment of immunity-related genes, vasculopathy, and fibrosis. Finally, the findings were integrated with current knowledge of SSc pathology, thereby identifying useful candidates for drug repurposing.

One of the main findings of the present study is that we could assign more than half of the eGenes (n = 134) to hallmarks of SSc pathogenesis. Interesting candidates were related to immune system processes, fibrosis, and vascular pathologies. Immune system processes highlighted eGenes like *CD247* or *BLK*, both of them previously associated with SSc and several ADs, such as RA or SLE (91,106,184). Regarding IFN-associated eGenes, we identified *IRF5* and the *IL12* receptors, *IL12RA* and *IL12RB,* which are well-established SSc risk loci, and are also associated with other ADs such as RA, SLE, and myositis, as described in chapter 3 (106,184). With regard to fibrosis, *TNXB* encodes a member of the tenascin family of extracellular matrix glycoproteins which is implicated in the regulation of the production and assembly of certain types of collagen (287). *TNXB* is also the main causative gene in Ehlers-Danlos syndrome, a connective tissue disorder characterized by altered skin elasticity, among other symptoms (288). The eGenes associated with vasculopathy or angiogenesis included *NOTCH4*, a non-classic HLA gene in the class II region that regulates *NOTCH1* and has previously been associated with SSc (177,289), and *CD151,* which is linked to vascular stability and neo-angiogenesis (290). Notch signaling has been demonstrated to contribute to collagen overproduction and fibroblast activation in SSc and fibrotic animal models, leading to a potential target pathway for future therapies in SSc fibrosis (291,292). Nevertheless, *NOTCH4*, unlike other Notch receptors, lacks detectable signaling capacity; acting through the inhibition of NOTCH1 signaling instead, when both receptors are expressed in the same cells (293). On the other hand, Notch signaling, and predominantly Notch1, promotes angiogenesis (294) being the development of abnormal angiogenesis one of the most important hallmarks of SSc. In this regard, previous studies in Notch1 $^{-/-}$ mutant mice models indicate an essential role of Notch1 signaling in the endothelium during vascular development (295). The lower expression of *NOTCH4* associated with SSc risk allele could lead to a lack of inhibition of NOTCH1 signaling, promoting abnormal angiogenesis and thus contributing to the vasculopathy and late loss of angiogenesis associated with SSc. Finally, regarding inflammatory processes, *C4A* and *C4B* are part of the complement system affected by active disease in a number of ADs (296). Interestingly, a recent study demonstrated the relevance of the copy number and resulting expression levels of *C4A* and *C4B*, as well as their contribution to sex-biased vulnerability in autoimmunity (297). In this regard, the eQTLs described in our study could be acting either as a proxy to *C4A*-*C4B* copy numbers or as an additional mechanism regulating the complex variation of complement genes.

Interestingly, we found 25 eGenes related to apoptosis processes. Previous genetic studies have indicated that apoptosis is an important mechanism of the disease, revealing the association of some genes, such as *DNASE1L3* or *TNFAIP3*, with a higher risk of SSc (101,298). We confirm here *DNASE1L3,* which plays an important role in DNA fragmentation during apoptosis (299), as an interesting candidate. Another eGene observed with a particular role in apoptosis was *BAK1*, which encodes for Bcl-2 antagonist or killer (BAK), one of the principal proapoptotic proteins of the mitochondrial pathway (300). Interestingly, a recent study showed that dermal fibroblasts derived from patients with SSc become particularly susceptible to apoptosis induced by mimetic drugs of proapoptotic protein Bcl-2 homology 3, a direct activator of BAK, reducing the fibrotic process (301). Thus, even though the specific pathogenic process of apoptosis in SSc is still unknown, our results support its role in SSc, which could be key to reversing fibrosis as part of the tissue regeneration process.

Regarding SSc specific eQTLs, the eGene *NCR3* highlights for its potential implication in SSc pathogenesis. *NCR3* encodes a natural cytotoxicity receptor implicated in the activation of NK cells. Interestingly, the protein encoded interacts with CD247, an important T cell receptor widely associated with SSc pathogenesis. In addition, Yau *et al* described a 33-kb region in the MHC III region overlapping the eQTL SNP of our study which regulates *NCR3* expression and contributes to increased autoimmune disease risk (302). Within these eQTLs found exclusively in SSc patients we identified an eQTL for *HLA-DRB1*, which is in fact the strongest candidate HLA gene predicting disease development (107).

It is noteworthy that 50% of the SSc eGenes associated with SSc hallmarks overlap with >1 group. This is not surprising, given that, for example, fibrosis, angiogenesis, and inflammation are closely linked, which demonstrates the complexity of the pathogenesis of SSc. Alternatively, there was significant enrichment of eQTLs in 19 tissues, most of them interpretable in the context of SSc, which affects tissues such as lungs, cardiac tissue, and esophagus (5). Surprisingly, we found that tyrhoid was the strongest enriched tissue. In line with this, the most common endocrine problem associated with SSc is thyroid disease, and several studies have documented that this disorder occurs in approximately 12% of SSc patients due to fibrosis of the thyroid gland (303). In addition, autoimmune regulator (*AIRE*) gene polymorphisms previously linked to SSc has been associated with autoinmune thyroiditis (304). Thus, our results emphasize the potential role of thyroid tissue in SSc pathogenesis.

A total of 24 transcription factor binding sites were stably enriched in best-model SSc-specific eQTLs. In this regard, the transcription factor *ELF1* (E74-like ETS transcription factor 1) deserves special mention, as it was also found to be differentially up-regulated in almost all 7 blood cell types assessed. *ELF1* belongs to the ETS family of transcription factors that regulate the expression of a wide range of genes and play an important role in immune cell development and function and in angiogenesis (305,306). This transcription factor activates the expression of several T cell genes. One of them is the gene encoding the ζ chain of the TCR, a molecule with a primary function in the transduction of intracellular signals that influence positive and negative selection of T cells upon TCR ligation (307). On the other hand, *ELF1* also plays an important role in B cells by cooperating with members of the activator protein 1 family of transcription factors to activate the 3′ immunoglobulin heavy-chain enhancer upon IgM stimulation, which could contribute to class-switch recombination (308). Of note, our enrichment analysis of TFBS should be interpreted with caution as the independence assumption of Fisher's exact test might not be fully met, since stepwise modeling does not necessarily generate independent loci for enrichment analysis.

Candidate eGenes identified here overlap with eQTL analyses performed in other ADs, further supporting our results and manifesting the shared genetic component of autoimmune disorders. Some eGenes, such as *BLK, GSDMB,* and *ORMDL3,* which have been described to be involved in RA (309), *KRT8P46, GSDMB,* and *ORMDL3*, involved in multiple sclerosis (310), *ANO9* and *BLK*, in SLE (311), and *GSMDA, GSDMB,* and *ORMDL3* in type 1 diabetes (312), were also significantly associated in our study.

Given the surprisingly high amount of candidate genes that warrant further studies, it is important to address the limits of this study. First, this study focused on bulk RNA-Seq and identified eQTLs present in the most abundant blood cell types. Although tools like CIBERSORT can successfully estimate the abundance of various cell types present, the number of samples needed to identify cell-specific eQTLs even in the most abundant cell types, using bulk RNA-Seq are still prohibitive (276). Second, although we highlight genes for which interpretation in the context of the disease is best understood in tissues other than blood, single-cell studies in SSc-affected tissues are needed to confirm and expand our findings. Last, we did not distinguish between the most common forms of SSc (limited cutaneous and diffuse cutaneous), nor did we analyze data on autoantibodies, as data were only available for a subset of the samples and would have severely diminished the sensitivity of our analysis.

The validation of the eQTLs identified from peripheral blood mononuclear cells (PBMCs) in other tissues as presented in the GTEx database, opens the way to cautiously use blood eQTLs as a proxy to detect eQTLs that most likely exert their main effect in tissues other than blood.

Interestingly, Beretta *et al* recently observed a strong enrichment of several IFN-related pathways in the first whole blood transcriptome profiling performed in a large cohort of SSc patients (146). Furthermore, a recent analysis of whole transcriptome expression in the skin of patients with early diffuse SSc revealed a high prevalence of both innate and adaptive immune cell activity (313). These results are concordant with the clear enrichment of immunity-related eGenes observed in our study and represent a support of the use of PBMC expression data as surrogate markers of organ disease.

To sum up, this is the first eQTL analysis performed in PBMCs of SSc patients, revealing that more than half of the eGenes detected were associated with the most important SSc hallmarks and highlighting the apoptotic process. Furthermore, we identified enriched motifs for transcription factors in SSc eQTLs that are differentially regulated in blood, skin, or lungs. Our results highlight the role of the clinical features and tissues involved in SSc, adding a new layer of complexity and contributing to a better understanding of SSc pathogenesis.

Chapter 5: Functional genomics in primary T cells and monocytes: linking genetic susceptibility loci with mechanisms influencing systemic sclerosis risk

5.1. Material and methods

5.1.1. Isolation of CD4$^+$ T cells and CD14$^+$ monocytes

Primary CD4$^+$ T cells and CD14$^+$ monocytes were collected from 10 SSc patients and 5 healthy individuals (cohort characteristics are described in **Supplementary Table S5.1**) with informed consent and with ethical approval. All SSc patients were diagnosed according to the American College of Rheumatology/European Alliance of Associations for Rheumatology 2013 criteria (37). PBMCs were isolated from 70 ml blood samples using Ficoll density gradient centrifugation. EasySep Human CD14$^+$ Positive Selection Kit (StemCell Technologies) was used to isolate CD14$^+$ cells from PBMCs and, subsequently, Easysep CD4$^+$ T Cell Isolation Kit (StemCell Technologies) was used to isolate CD4$^+$ T cells from the remaining PBMCs, according to the manufacturer's instructions.

5.1.2. Promoter capture Hi-C probe design

First, gene annotations for 18,755 protein coding genes were extracted from Ensembl's genebuild database version 97 GRCh38. The TSS for each gene was located using the first base of the gene coordinates with respect to gene orientation. Capture regions were identified by mapping the TSS coordinates to the *in silico* digested (Arima Hi-C) genome and extracting the fragments containing the TSS coordinates as well as the fragments directly upstream and downstream. Therefore, each TSS is represented by a total of 3 contiguous restriction fragments. The average length of the 3 consecutive restriction fragments for each TSS is 786bp and the median is 927bp, with a range of 54-4174bp. If the length of the restriction fragment is no more than 700bp, the entire restriction fragment is covered by the probes. If the length is more than 700bp, then the middle part will not be in the probe design. Using this final set of restriction fragments, a BED file was prepared for input into the SureDesign (Agilent) probe design web tool. Probes were designed using a 1X tiling approach, with moderate repeat masking and maximum performance boosting optimized for the SureSelect XT HS and XT Low Input capture workflows.

5.1.3. Capture Hi-C library generation

5-10 million isolated CD4+ T cells and CD14+ monocytes were crosslinked for 10 min in 1% formaldehyde and agitated at room temperature, the reaction was then quenched with ice cold 0.125 M glycine for 5 min. Crosslinked cells were washed in ice-cold PBS, and the supernatant was discarded, the pellets were then stored at -80ºC.

Each Hi-C library was prepared from fixed cells following the Arima HiC kit (Arima Genomics) and the KAPA HyperPrep kit (KAPA Biosystems), following the manufacturer's protocol. Briefly, crosslinked cells were lysed, and DNA was digested by two different restriction enzymes and ligated. The ligated DNA was reverse-crosslinked and fragmented by sonication (Covaris S220), followed by DNA size selection, biotin enrichment, adaptor ligation, and final library amplification. A quality control step to determine the number of PCR cycles needed for library amplification was performed for each library using the NEBNext Library Quant kit (NEBNext). Final quality and quantity were assessed by the Bioanalyzer 2100 (Agilent) and Qubit (Thermo Fisher) systems.

Hi-C samples were then hybridized with the SureSelect custom capture library by following Agilent SureSelectXT HS reagents and protocols. Briefly, target regions are hybridized with designed probes, followed by capture enrichment of these regions through streptavidin pulldown using Dynabeads MyOne Streptavidin T1 (ThermoFisher, ref: 65601). Post-capture amplification was carried out using nine PCR cycles. Final quality and quantity were assessed by the Bioanalyzer 2100 and Qubit systems.

5.1.4. Promoter capture Hi-C (pCHi-C) sequencing and processing

Sequencing for 30 prepared pCHi-C libraries was performed using five lanes of Illumina NovaSeq S4 flow cell on Illumina NovaSeq6000, generating 150 bp paired-end reads, and leading to an average of 500 million reads per sample (Novogene Company LTD). Sequencing data was filtered and the adapters were removed using fastp v0.19.4 (314). Subsequent mapping with GRCh38 and filtering was performed with HiCUP v0.7.4 (315) and bowtie2 v2.3.2 (316), taking as maximum and minimum di-tag lengths 700 and 100, respectively. Only intrachromosomal interactions were included in the analysis, and off-target di-tags where neither end mapped to a targeted fragment were removed (statistics in **Supplementary Table S5.2**). Significant chromatin interactions were identified using CHiCAGO v1.13.1 (317) using a threshold of CHiCAGO score > 5 in different conditions; cell type: $CD4^+$ T cells (n=15 biological replicates) and $CD14^+$ monocytes (n=15); cell type and disease state: $CD4^+$ T cells from SSc patients (n=10) and healthy controls (n=5), and $CD14^+$ monocytes from SSc patients (n=10) and healthy controls (n=5). PCA was performed in each cell type in order to detect potential biases and the two first PCs were plotted using R 3.6.1 (**Supplementary Figure S5.1**). Libraries generated using the Arima Hi-C protocol have different characteristics from the original CHi-C protocol, for this reason, several changes were made to the analysis. The restriction fragments were binned for the analysis as follows: every 20 consecutive restriction fragments were binned into one bin; if a baited region was encountered, a separate bin was created for the baited region that includes the 3 captured fragments and the fragment before and after for a total of 5 fragments per region captured. Consecutive baits were merged into a single bait. This design led to a baitmap of 17,313 baited regions, capturing 18,630 promoters. Default settings were used for the CHiCAGO pipeline, except for design file parameters minFragLen and maxFragLen, set to 140 and 20000, respectively. The weights of the model used by CHiCAGO in the defined conditions were re-estimated using the function "fitDistCurve.R" included in CHiCAGO package after running the pipeline on a merged file including the total 30 samples.

In order to detect enrichment of features in the interactions obtained with CHiCAGO for each cell type, narrowPeak bed files of H3K4me3 and H3K27ac were obtained as follows: H3K4me3 (ENCODE; ENCFF828YVL) and H3K27ac (ENCODE; ENCFF790PVU) in primary CD4$^+$ naive T cells; H3K4me3 (ENCODE; ENCFF651GXK) and H3K27ac (ENCODE; ENCFF471QGU) in primary CD14$^+$ monocytes. The "peakEnrichment4Features" function from the CHiCAGO package was used to detect enrichment of each feature in pCHi-C data. Finally, Chicdiff v0.6 (318) R package was used to detect differential interactions between different conditions: CD4$^+$ T cells vs CD14$^+$ monocytes; SSc patients vs healthy controls CD4$^+$ T cells; and SSc patients vs healthy controls CD14$^+$ monocytes. Code was modified to include age, sex, and disease status (only in cell type comparison) as covariates. For each comparison, only those interactions with CHiCAGO score > 5 in at least one condition were included in the differential analysis. Differential interactions with a weighted adjusted p-value < 0.05 were identified as significant. Spearman's rank-order correlation was performed to test the correlation of $\log_2$ fold change values in differential interactions between patients and controls in CD4$^+$ T cells and CD14$^+$ monocytes.

5.1.5. RNA-seq library generation

A total of 0.5 million purified cells was resuspended in 700 uL Qiazol lysis reagent (QIAGEN) to isolate RNA, then 140 uL of chloroform was added.

After centrifugation at 12000 x g for 15 min, approximately 350 uL of the upper layer containing the RNA was transferred and mixed with 525 uL of 100% ethanol. RNA isolation was continued from this point using the RNeasy microkit (QIAGEN) reagents and protocol. Final quantity was assessed by Qubit fluorimeter. Libraries for RNA-seq were prepared using Illumina Truseq Stranded Total RNA reagents and protocol, except for CD4$^+$ and CD14$^+$ samples from Control 1, for which library preparation failed. Library quality and quantity was assessed by Bioanalyzer. The 28 libraries generated were sequenced using three lanes on Illumina HiSeq4000, generating 75 bp paired-end reads, and leading to an average of 30 million reads per sample (Genomics Facility, University of Manchester).

5.1.6. RNA-seq data processing

RNA-seq reads were quality trimmed and adapters were removed using fastp v0.19.4 (314). Reads were then mapped using STAR v2.7.3a (269) on the GRCh38 genome with GENCODE annotation v32. Reads were de-duplicated with Picard tools v2.22.2 (http://broadinstitute.github.io/picard/) and then counted using HTSeq v0.12.3 (319) (**Supplementary Table S5.3**). Final count matrices were analysed in R 3.6.1, using edgeR v3.28.1 (277) to perform normalization and differential expression analysis. Three differential expression comparisons were performed: CD4$^+$ T cells vs CD14$^+$ monocytes; SSc patients vs healthy controls in CD4$^+$ T cells; and SSc patients vs healthy controls in CD14$^+$ monocytes. Age, sex, and disease status (only in cell type comparison) were used as covariates. Only those genes with a mean of more than 1 $\log_2$ transformed counts per million (CPM) across all samples were included in further analysis, for a total of 11,221 coding genes. Differentially expressed genes were called with an adjusted *p*-value of 0.1 (FDR 10%). Due to the extreme differences in expression regarding the CD4$^+$ vs CD14$^+$ comparison, only those with an absolute value of $\log_2$ fold change ($|\log_2 FC|$) > 2 and FDR < 5% were taken in further analysis. Functional enrichment analyses were performed with g:Profiler (320) with default settings, taking gene ontology and biological pathways as data sources.

5.1.7. Linking differential expression and differential interactions in CD4⁺ T cells vs CD14⁺ monocytes

Genes corresponding with the promoter end of significant differential interactions observed between $CD4^+$ and $CD14^+$ cells were overlapped with those differentially expressed. One-sided Fisher's exact test was performed in order to calculate the enrichment of genes with differential interactions in those differentially expressed. In this set of overlapping genes, Spearman's rank-order correlation was performed to test the correlation of $\log_2$ fold change values in differential interactions and differential expression. The $\log_2$ fold change value for each gene with differential interactions was obtained as the median $\log_2$ fold change value of all interactions corresponding to a specific promoter. Finally, in order to test the distribution of $\log_2$ fold change values, a binomial exact test was performed on a subset of overlapping genes obtained adding a more stringent cutoff, including only those differentially interacting genes with an absolute value of median $\log_2 FC$ > 2 for each gene. Functional enrichment analyses were performed with g:Profiler (320) with default settings, taking gene ontology and biological pathways as data sources.

5.1.8. Defining SSc-associated GWAS loci

All independent non-MHC SSc-associated signals were selected from the largest meta-GWAS performed to date (106). We defined 23 regions based on LD data and SNP proximity from the total of 27 independent signals described in this GWAS. First, we took all SNPs associated at the genome-wide significant level (*p*-value < $5x10^{-8}$) (427 SNPs) as well as those SNPs in high

LD with them ($r^2 > 0.8$) from the meta-GWAS datasets (106) using PLINK v1.9 (170), resulting in a total of 1,505 SNPs. To facilitate pCHi-C analysis, independent signals corresponding to the same locus were grouped. These grouped independent signals interacted with the same promoters, and did not represent a change if they were taken as separated loci. The window ranges and total number of SNPs in each of the 23 final loci are specified in **Supplementary Table S5.4**.

5.1.9. Defining enhancers and TADs in CD4⁺ T cells and CD14⁺ monocytes

In order to define enhancer regions, chromHMM v1.22 annotations (321) from 9 CD4$^+$ T cells (BSS00183, BSS00185, BSS00186, BSS00188, BSS00189, BSS00190, BSS00191, BSS00192 and BSS00274) and 4 CD14$^+$ monocytes (BSS00178, BSS00179, BSS00180, and BSS00181) were downloaded from the EpiMap project (322). For each cell type, enhancer regions were defined as those with state number from chromHMM corresponding to enhancer activity (7, 8, 9 ,10, 11, and 15) present in at least one sample. Topologically associated domains (TADs) definition for CD4$^+$ T cells and CD14$^+$ monocytes was obtained from Javierre *et al* (166).

5.1.10. Overlap among pCHi-C, SSc-associated GWAS loci, and enhancer regions

In order to prioritize certain interactions observed in pCHi-C data of particular interest in SSc GWAS loci, the SNP set previously defined in “Defining SSc-associated GWAS loci'' was overlapped with enhancer regions of each cell type using the GenomicRanges (323) package implemented in R 3.6.1 (**Supplementary Table S5.4**). This new SNP set was then overlapped with the promoter interacting regions (PIRs) of significant pCHi-C interactions, defining candidate interacting genes as those in which their PIR overlaps with our significant SSc SNP set and enhancer regions. One-sided

Fisher's exact test was performed in order to calculate the enrichment of the SNP set in CD4$^+$ T cells and CD14$^+$ monocytes enhancer regions. A test for equality of proportions or two proportion z-test ("prop.test" in R 3.6.1) was performed to calculate the enrichment of SNPs overlapping enhancer regions between cell types, correcting by total number of base pairs covered by enhancer regions for each cell type. For each candidate interacting gene, the median of $\log_2$FC values and weighted adjusted *p*-values was calculated taking all differential interactions with the PIR overlapping our SSc SNP set and enhancer regions. Functional enrichment analyses were performed for the sets of interacting genes observed in CD4$^+$ T cells and CD14$^+$ monocytes with g:Profiler (320) using default settings, taking gene ontology and biological pathways as data sources.

5.1.11. Visualization tools

The WashU Epigenome browser (324) was used to plot pCHi-C interactions, enhancer regions defined by chromHMM and H3K27ac peaks (from EpiMap samples previously defined), and TADs (from Javierre *et al* (166)) in CD4$^+$ T cells and CD14$^+$ monocytes.

5.1.12. Drug target analysis

In order to assess if genes interacting with SSc GWAS loci in CD4$^+$ T cells and CD14$^+$ monocytes presented potential drug targets that could be repurposed for their use in SSc, those genes interacting with PIR overlapping SSc-associated SNPs and enhancer regions, were used to model a PPI network using STRING v11 (203) with the highest interaction confidence score (>0.9), calculated as a combined probability from different evidences of interactions corrected for the probability of observing a random interaction. Protein products from these genes and those in direct PPI with them were used to query the OpenTargets Platform (235) for drug targets. Additionally, the same platform and the Drugbank database (236) were searched for information on clinical studies of drug targets of interest in SSc. Only drug targets with at least completed phase III clinical trials in SSc and/or similar immune-mediated diseases were included in the results.

5.2. Results

We generated pCHi-C data for CD4$^+$ T cells and CD14$^+$ monocytes from 10 SSc patients and 5 healthy controls. CHiCAGO was used to identify significant promoter interactions (CHiCAGO score > 5) for each cell type and disease condition (**Supplementary Table S5.5**) and Chicdiff was used to identify differential interactions between cell types, and between disease conditions for each cell type. A total of 81,624 and 74,853 significant interactions corresponding to 8,193 and 7,024 captured promoters were identified in CD4$^+$ T cells and CD14$^+$ monocytes, respectively. In addition, 71,213 significant differential interactions (weighted adjusted p-value < 0.05) corresponding to 8,223 captured promoters were obtained in the comparison between cell types. Through integration with published ChIP-seq data, we found that PIRs were enriched in H3K27ac and H3K4me3 histone marks from primary CD4$^+$ naive T cells and CD14$^+$ monocytes (**Figure 5.1**), suggesting that promoters mostly interact with regulatory active regions such as enhancers.

5.2.1. Differential interactions and expression between SSc patients and healthy controls

One of the main aims of this study was to identify specific interactions that could be present in SSc patients but not in healthy controls, or vice versa, and thus, identify specific genes interacting with enhancer regions and SSc-associated loci that could be of interest in SSc pathology. With this in mind, a

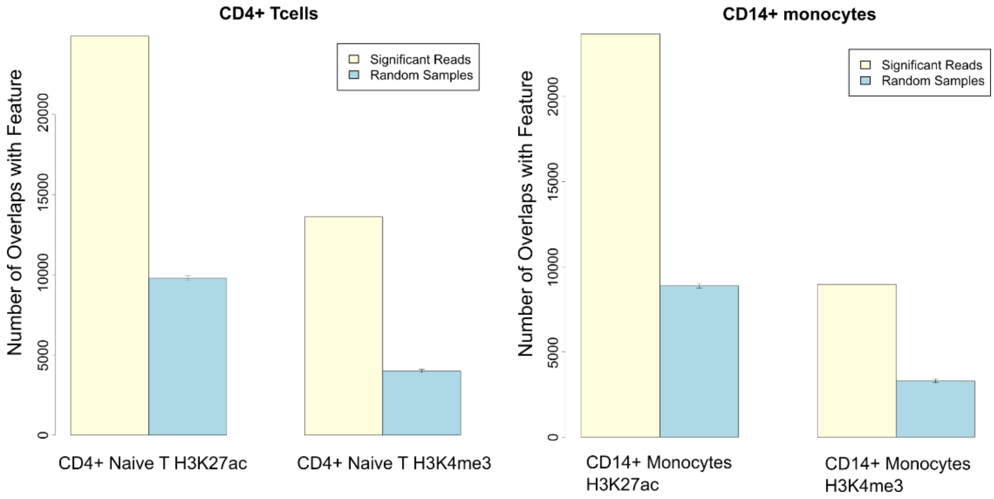


Figure 5.1. Enrichment of features within promoter interacting regions (PIRs) of pCHi-C interactions. Peak locations of H3K4me3 and H3K27ac (ENCODE) in primary CD4$^+$ naive T cells and CD14$^+$ monocytes were tested against promoter interacting regions (PIRs) of interactions using the "peakEnrichment4Features" function of CHiCAGO package. The graphs show the number of overlaps with the feature in the interaction data (yellow) vs the mean number of overlaps in 100 sampled interactions from the non-significant pool (blue). Error bars correspond to the 95% confidence interval.

total of 4,858 significant differential interactions (weighted adjusted *p*-value < 0.05) were identified between SSc patients and healthy controls in CD4$^+$ T cells, corresponding to 1,526 captured promoters. Despite this, any significant differential interactions in CD14$^+$ monocytes were detected. Further analysis of the differential interactions in CD4$^+$ T cells revealed that their presence could be due to a possible bias as all of them showed scores just above significance threshold (median weighted adjusted *p*-value = 2.2x10^{-2}) as compared with differential interactions between cell types (median weighted adjusted *p*-value = 2.16x10^{-10}), and PCA showed some potential batch effect particularly in controls from CD4$^+$ T cells (**Supplementary Figure S5.1**). In addition, taking those genes with differential interactions in CD4$^+$ T cells, gene set enrichment analysis showed no particular pathway of interest in SSc pathology. Nevertheless, a significant positive correlation was found in log$_2$FC values from SSc vs controls differential interactions between both cell types (Spearman's rank correlation *p*-value = 5.22x10^{-173}, rho = 0.35) (**Supplementary Figure S5.2**). On the other hand, none of the 23 loci (defined in "Material and Methods") corresponding to SSc GWAS associated regions showed significant differences at the interaction level between patients and controls. A lack of statistical power could be the main reason of these negative results, probably caused by a sample size not large enough to identify slight differences. Besides, a total of 62 and 63 differentially expressed genes (FDR < 0.05) were identified between patients and controls in CD4^{+} T cells and CD14^{+} monocytes, respectively (**Supplementary Tables S5.6** and **S5.7**). In the case of CD4^{+} T cells, we observed significant enrichment in terms related with immune response such as "positive regulation of immune system process" or "leukocyte activation" (**Supplementary Table S5.8**). However, we could not identify any functional enrichment regarding the 63 genes differentially expressed in CD14^{+} monocytes.

5.2.2. Linking differential expression and differential interactions in CD4^{+} T cells vs CD14^{+} monocytes

Due to the inconclusive results observed in the comparison between SSc patients and healthy controls, we decided to look at general differences at the interaction and expression levels between cell types, without taking disease state into account. On this subject, a total of 19,125 protein coding genes were analyzed in the RNA-seq data, of which 9,795 were identified as differentially expressed between CD4^{+} T cells and CD14^{+} monocytes. Subsequently, 2,257 strongly differentially expressed genes were obtained (FDR < 0.05, $|\log_2 FC| > 2$), of which 919 and 1,338 genes were overexpressed in CD4^{+} T cells and CD14^{+} monocytes, respectively. Overrepresentation analyses showed that each group of genes is, as expected, significantly enriched in terms related with T cells and monocytes specific pathways, including gene ontology terms such as "T cell activation" and "T cell differentiation" in CD4$^+$ T cells, and "leukocyte activation" in CD14$^+$ monocytes (**Supplementary Tables S5.9** and **S5.10**). In addition, we observed that differentially expressed genes are in fact significantly enriched in differentially interacting genes (fisher exact test *p*-value = 3.54x10^{-37}, OR = 1.77). Furthermore, from the total of 1,209 differentially expressed genes overlapping differentially interacting genes, we observed that genes overexpressed in a specific cell type significantly correlated with increased number of interactions in that cell type, and vice versa (Spearman's rank correlation *p*-value = 1.04x10^{-197}, rho = 0.73).

Finally, we applied a more stringent cutoff in differentially interacting genes ($|\log_2 FC| > 2$), leading to a total of 97 differentially expressed genes overlapping differentially interacting genes. In this subset, only 2 of the 97 genes were not distributed as expected; whilst 23 and 72 genes were overexpressed and presented an increased number of interactions in CD4$^+$ T cells and CD14$^+$ monocytes, respectively (exact binomial test *p*-value =6.01x10^{-26}, probability of success = 98%) (**Supplementary Figure S5.3**). In this regard, overrepresentation analysis performed in the subset of 23 genes (CD4$^+$ T cells) showed an enrichment in T cell related terms (**Supplementary Table S5.11**), including important genes in T cell differentiation and activation such as *BCL11B* or *LEF1*. On the other hand, the subset of 72 genes corresponding to CD14$^+$ monocytes showed an enrichment in monocyte related and other terms (**Supplementary Table S5.12**), including important genes in macrophage differentiation and modulation of monocyte inflammatory response such as *PADI2*, *S100A8* and *CXCL8*. Thus, our results demonstrate the importance of using different cell types to define promoter interactions and how they are linked with the expression of important genes for those cell types. In this regard, we decided to define significant interactions that linked SSc-associated loci with promoters in CD4$^+$ T cells and CD14$^+$ monocytes.

5.2.3. SSc-associated loci and CD4$^+$ / CD14$^+$ promoter interactions

To identify new potential target genes associated to SSc, as well as the potential implication of different cell types in those associations, we performed a multi-omic approach overlapping 23 regions defined based on the most powerful SSc meta-GWAS performed to date (106) with enhancer regions and our pCHi-C data. From the total of 1,505 SNPs, including those associated with SSc at the genome-wide significance level (p-value < 5x10^{-8}) and their proxies ($r^2 > 0.8$), 445 (29.6%) and 284 (18.9%) overlapped with enhancer regions from CD4$^+$ T cells and CD14$^+$ monocytes, respectively. As expected, the GWAS SNP set was significantly enriched in both CD4$^+$ and CD14$^+$ enhancer regions (one-sided Fisher's exact test p-value = 5.91x10^{-133}, OR = 4.86 in CD4$^+$ T cells; p-value = 1.63x10^{-56}, OR = 3.27 in CD14$^+$ monocytes). In addition, significant differences in the number of SNPs overlapping CD4$^+$ and CD14$^+$ enhancer regions were identified (two proportion z-test p-value = 0.001), observing a stronger overlap with CD4$^+$ T cell enhancer regions as compared with those from CD14$^+$ monocytes. These GWAS SNPs within enhancer regions were overlapped with PIRs from pCHi-C, obtaining a total of 398 and 109 significant interactions in CD4$^+$ T cells and CD14$^+$ monocytes, respectively (**Supplementary Table S5.4**). The promoter ends of those interactions correspond to 46 genes, with a total of 40 and 27 interacting genes in CD4$^+$ T cells and CD14$^+$ monocytes, respectively (**Table 5.1**).

Table 5.1. pCHi-C target genes for the 23 systemic sclerosis associated regions in CD4+ T cells and CD14+ monocytes.

Chr	Bp (start - end)[1]	GWAS Locus[2]	pCHi-C target genes[3]		CD4+ vs CD14+	
			CD4+ T cells	CD14+ monocytes	Differential interactions	Differential expression
1	67326053 - 67448804	*IL12RB2*				
1	167445635 - 167465040	*CD247*	***CD247***, *CREG1*		***CD247***, *CREG1*	***CD247***, *CREG1*
1	173337507 - 173391947	*TNFSF4- LOC100506023 -PRDX6*				
2	190642047 - 190698201	*NAB1*	*MFSD6, NEMP2*	*MFSD6, NEMP2, HIBCH;INPP1*	*MFSD6, NEMP2, HIBCH;INPP1*	*MFSD6, NEMP2, HIBCH, INPP1*
2	191035723 - 191108308	*STAT4*	***STAT4***, *NABP1*		***STAT4***, *NABP1*	***STAT4***, *NABP1*
3	58084620 - 58482701	*FLNB -DNASE1L3-PXK*	*RPP14, KCTD6*	*RPP14, KCTD6*	*RPP14, KCTD6*	*KCTD6*
3	119384733 - 119546340	*POGLUT1-TIMMDC1-CD80- ARHGAP31*	*TMEM39A;POGLUT1*		*TMEM39A;POGLUT1*	*TMEM39A, POGLUT1*
3	160002484 - 160030580	*IL12A*		*SMC4;IFT80*	*SMC4;IFT80*	*SMC4, IFT80*
4	960523 - 990021	*DGKQ*	*GAK;TMEM175, FGFRL1*	*GAK;TMEM175, FGFRL1*	*FGFRL1*	*GAK, TMEM175 FGFRL1*
4	102477892 - 102615256	*NFKB1*	*SLC39A8*, ***NFKB1***, *UBE2D3;CISD2, SLC9B1, BDH2*	*SLC39A8, UBE2D3;CISD2, BDH2*	*SLC39A8*, ***NFKB1***, *UBE2D3;CISD2, SLC9B1, BDH2*	*SLC39A8, UBE2D3, CISD2, BDH2*
5	151064651 - 151080486	*TNIP1*				

Genes classically associated with systemic sclerosis through proximity to GWAS loci are highlighted in bold.

[1] Bp in GRCh38 (hg 38) assembly

[2] Locus as defined by López-Isac et al. *Nat Genet.* 2019

[3] Genes corresponding with promoter interacting regions overlapping enhancer regions and systemic sclerosis GWAS SNPs

Table 5.1. pCHi-C target genes for the 23 systemic sclerosis associated regions in CD4+ T cells and CD14+ monocytes (continuation)

Chr	Bp (start - end)[1]	GWAS Locus[2]	pCHi-C target genes[3]		CD4⁺ vs CD14⁺	
			CD4⁺ T cells	CD14⁺ monocytes	Differential interactions	Differential expression
6	106181815 - 106339294	*ATG5*				
7	128933913 - 129095960	*IRF5-TNPO3*				
8	11474517 - 11544554	*FAM167A-BLK*				
8	60638547 - 60664239	*RAB2A-CHD7*	*ASPH, SDCBP, CHD7*		*ASPH, SDCBP*	*ASPH, SDCBP, CHD7*
11	554659 - 619789	*CDHR5 -IRF7*				
11	2311894 - 2363262	*TSPAN32,CD81-AS1*	*TSSC4*		*TSSC4*	*TSSC4*
11	118704617 - 118875175	*DDX6*	*CXCR5, UPK2,* ***DDX6****, IFT46;ARCN1*	*CXCR5*	*CXCR5, UPK2,* ***DDX6***	*CXCR5,* ***DDX6****, ARCN1*
15	74739180 - 75148328	*CSK*	***CSK****, CLK3, ULK3, SCAMP2, MPI, FAM219B, COX5A*	***CSK****, CLK3, ULK3, SCAMP2, MPI, FAM219B, COX5A, C15orf39*	***CSK****, ULK3, SCAMP2, MPI, FAM219B, COX5A, C15orf39*	***CSK****, CLK3, ULK3, SCAMP2, MPI, FAM219B, COX5A, C15orf39*
16	85932852 - 85979945	*IRF8*		***IRF8***	***IRF8***	***IRF8***
17	39747478 - 39933464	*IKZF3-GSDMB*	***IKZF3****, ERBB2, PSMD3*	*ERBB2*	***IKZF3****, ERBB2, PSMD3*	***IKZF3****, ERBB2, PSMD3*
17	75193533 - 75279345	*NUP85-GRB2*				
19	18068862 - 18093031	*IL12RB1*	*PIK3R2, RAB3A*	*RAB3A*	*PIK3R2, RAB3A*	

Genes classically associated with systemic sclerosis through proximity to GWAS loci are highlighted in bold.

[1] Bp in GRCh38 (hg 38) assembly

[2] Locus as defined by López-Isac *et al.* Nat Genet. 2019

[3] Genes corresponding with promoter interacting regions overlapping enhancer regions and systemic sclerosis GWAS SNPs

The physical interaction maps presented here identify 39 new potential candidate genes and confirm 7 genes which have been associated by classical GWAS methods using proximity. Differential expression and differential interaction data for each of the 46 genes and baited promoters, respectively, are available in **Supplementary Tables S5.13** and **S5.14**. Interestingly, some SSc confirmed genes such as *IRF8*, *STAT4*, or *CD247*, showed cell type specific interactions (**Figures 5.2-4**). The *IRF8* locus (**Figure 5.2**) provides a good example in which interactions between SNPs overlapping enhancer regions (represented by H3K27ac mark peaks) and *IRF8* promoter are exclusively found in one cell type and are associated to differential gene expression between cells, in this case corresponding to CD14+ monocytes, that showed a much higher expression of *IRF8* ($\log_2$FC = -4.47, FDR = 3.11x10^{-72}). In the case of *STAT4* (**Figure 5.3**), we did not detect significant interactions between SNPs overlapping enhancer regions and gene promoter in CD14$^+$ monocytes. On the other hand, significant interactions with the *STAT4* promoter were identified in CD4$^+$ T cells, corresponding with a TAD specific for CD4$^+$ T cells that is not found in monocytes. In addition, *STAT4* showed a significantly higher expression in CD4$^+$ T cells as compared with CD14$^+$ monocytes ($\log_2$FC = 7.05, FDR = 1x10^{-304}). We also found significant interactions in this same locus with the promoter of *NABP1* in CD4$^+$ T cells, located 600 kb downstream, crossing TAD boundaries in both cell types. Here, we identified a significant differential expression of *NABP1* ($\log_2$FC = -0.31, FDR = 2.86x10^{-3}), slightly overexpressed in CD14$^+$ monocytes. Cell type specific interactions were also observed for the *CD247* locus (**Figure 5.4**), in which significant interactions between SNPs and *CD247* promoter were evident only in CD4$^+$ T cells, with an increased expression of this gene in CD4$^+$ T cells as compared with CD14$^+$ monocytes ($\log_2$FC = 7.49, FDR = 3.99x10^{-210}). Furthermore, significant interactions with the promoters of *CREG1* were identified in the same locus in CD4$^+$ T cells. In this regard, we observed an overexpression of *CREG1* in CD14$^+$ monocytes ($\log_2$FC = -3.68, FDR = 1.32x10^{-325}).

On the other hand, we identified new potential candidate genes interacting with SSc-associated SNPs by GWAS historically associated with the closest gene. One of these examples correspond to the *DDX6* locus (**Figure 5.5**) in which we found significant interactions between SNPs overlapping enhancer regions and not only *DDX6*, but also other potential candidate genes including *CXCR5*, *UPK2*, and *IFT46;ARCN1* (shared capture bait) promoters in CD4$^+$ T cells. In the case of CD14$^+$ monocytes only a significant interaction with the *CXCR5* promoter was found. All of these interactions are intra-TAD, except for the one including the *IFT46;ARCN1* promoters. In addition, we observed a significantly higher gene expression of *CXCR5* ($\log_2$FC = 3.21, FDR = 1.05x10^{-9}) and *DDX6* ($\log_2$FC = 1.14, FDR = 2.38x10^{-83}) in CD4$^+$ T cells, while *ARCN1* showed a slight overexpression in CD14$^+$ monocytes ($\log_2$FC = -0.21, FDR = 3.69x10^{-3}). *IFT46* was not significantly differentially expressed ($\log_2$FC = 0.25, FDR = 1.08x10^{-1}) (*UPK2* differential expression could not be tested due to low expression levels).

In order to identify what pathways could be driving disease in the two different cell types, we performed a functional enrichment analysis including the genes interacting with SSc-associated loci for each cell type. In CD4$^+$ T cells, the set of 40 interacting genes observed showed enrichment in virus response and pancreatic carcinoma (**Supplementary Table S5.15**). In accordance with this, a higher incidence of cancer in SSc patients compared with the general population has been suggested in several studies (325). On the other hand, the set of 27 interacting genes observed in CD14$^+$ monocytes showed enrichment in tyrosine kinase activity (**Supplementary Table S5.16**), which plays an important role in fibrosis, and has been related with SSc pathogenesis, being tyrosine kinase inhibitors one of the most promising antifibrotic therapies for SSc and other fibrotic diseases (326).

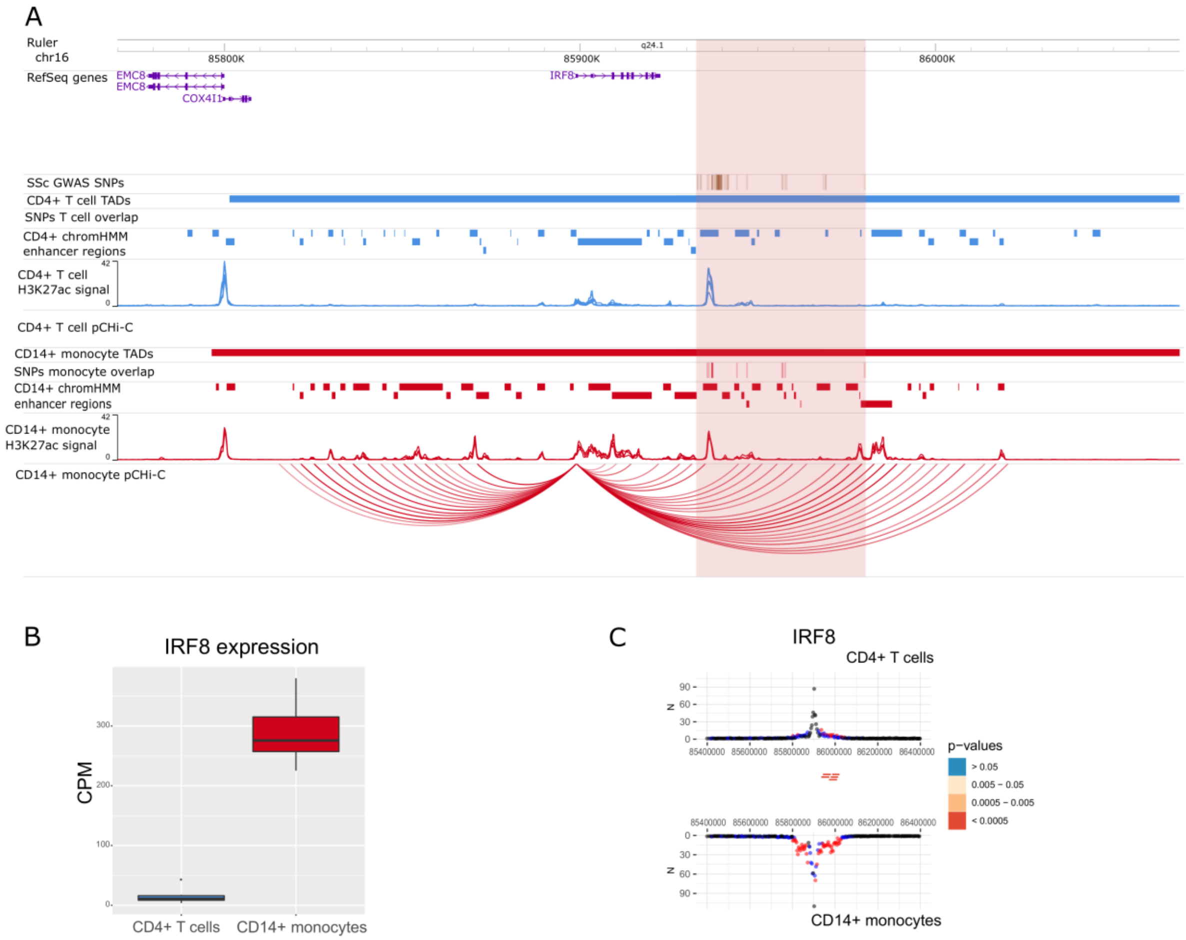


Figure 5.2. Promoter capture Hi-C (pCHi-C) interactions and gene expression in the *IRF8* locus. (**A**) Genomic coordinates (GRCh38) are shown at the top of the panel. The tracks include RefSeq genes (NCBI), systemic sclerosis (SSc)-associated SNPs from López-Isac *et al* (106) and their proxies (r^2>0.8), topologically associating domains (TADs) (shown as bars), SNPs overlapping promoter interacting regions (PIRs) and enhancer regions, enhancer regions as defined by chromHMM, H3K27ac signal, and pCHi-C significant interactions (CHiCAGO score > 5) (shown as arcs) in CD4$^+$ T cells (blue) and CD14$^+$ monocytes (red). The highlighted region in red includes all the SSc-associated SNPs LD block. (**B**) Gene expression level of *IRF8* from CD4$^+$ T cells and CD14$^+$ monocytes in counts per million (CPM). (**C**) Chicdiff bait profiles were generated for the *IRF8* gene. The plot shows the raw read counts versus linear distance from the bait fragment as mirror images for CD4$^+$ T cells and CD14$^+$ monocytes. Other-end interacting fragments are pooled and color-coded by their adjusted weighted *p*-value. Significant differentially interacting regions detected by Chicdiff overlapping SSc-associated SNPs and enhancer regions are depicted as red blocks.

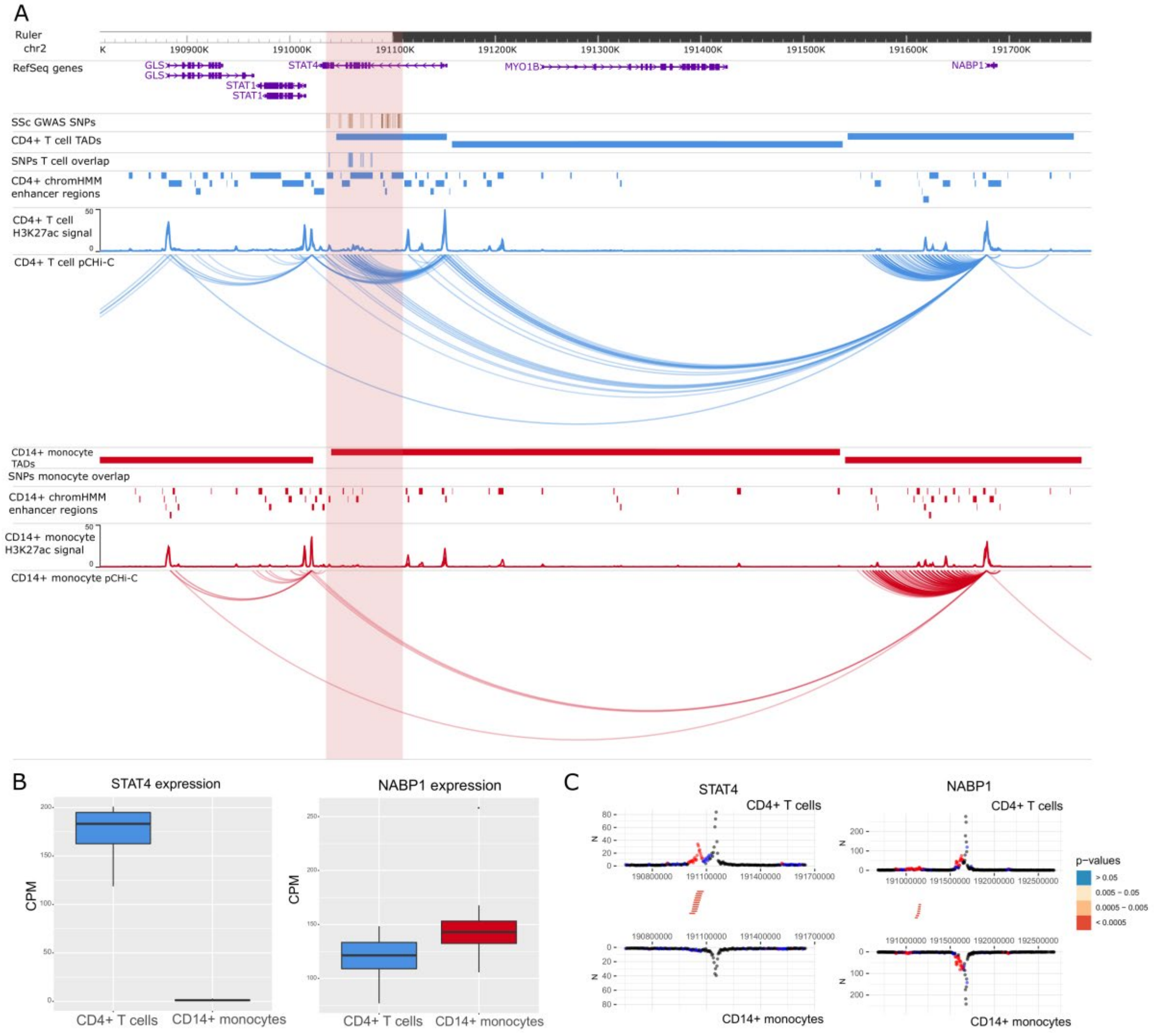


Figure 5.3. Promoter capture Hi-C (pCHi-C) interactions and gene expression in the *STAT4* GWAS locus. (**A**) Genomic coordinates (GRCh38) are shown at the top of the panel. The tracks include RefSeq genes (NCBI), systemic sclerosis (SSc)-associated SNPs from López-Isac *et al* (106) and their proxies (r^2>0.8), topologically associating domains (TADs) (shown as bars), SNPs overlapping promoter interacting regions (PIRs) and enhancer regions, enhancer regions as defined by chromHMM, H3K27ac signal, and pCHi-C significant interactions (CHiCAGO score > 5) (shown as arcs) in CD4$^+$ T cells (blue) and CD14$^+$ monocytes (red). The highlighted region in red includes all the SSc-associated SNPs LD block. (**B**) Gene expression level of *STAT4* and *NABP1* from CD4$^+$ T cells and CD14$^+$ monocytes in counts per million (CPM). (**C**) Chicdiff bait profiles were generated for the *STAT4* and *NABP1* genes. Plots show the raw read counts versus linear distance from the bait fragment as mirror images for CD4$^+$ T cells and CD14$^+$ monocytes. Other-end interacting fragments are pooled and color-coded by their adjusted weighted *p*-value. Significant differentially interacting regions detected by Chicdiff overlapping SSc-associated SNPs and enhancer regions are depicted as red blocks.

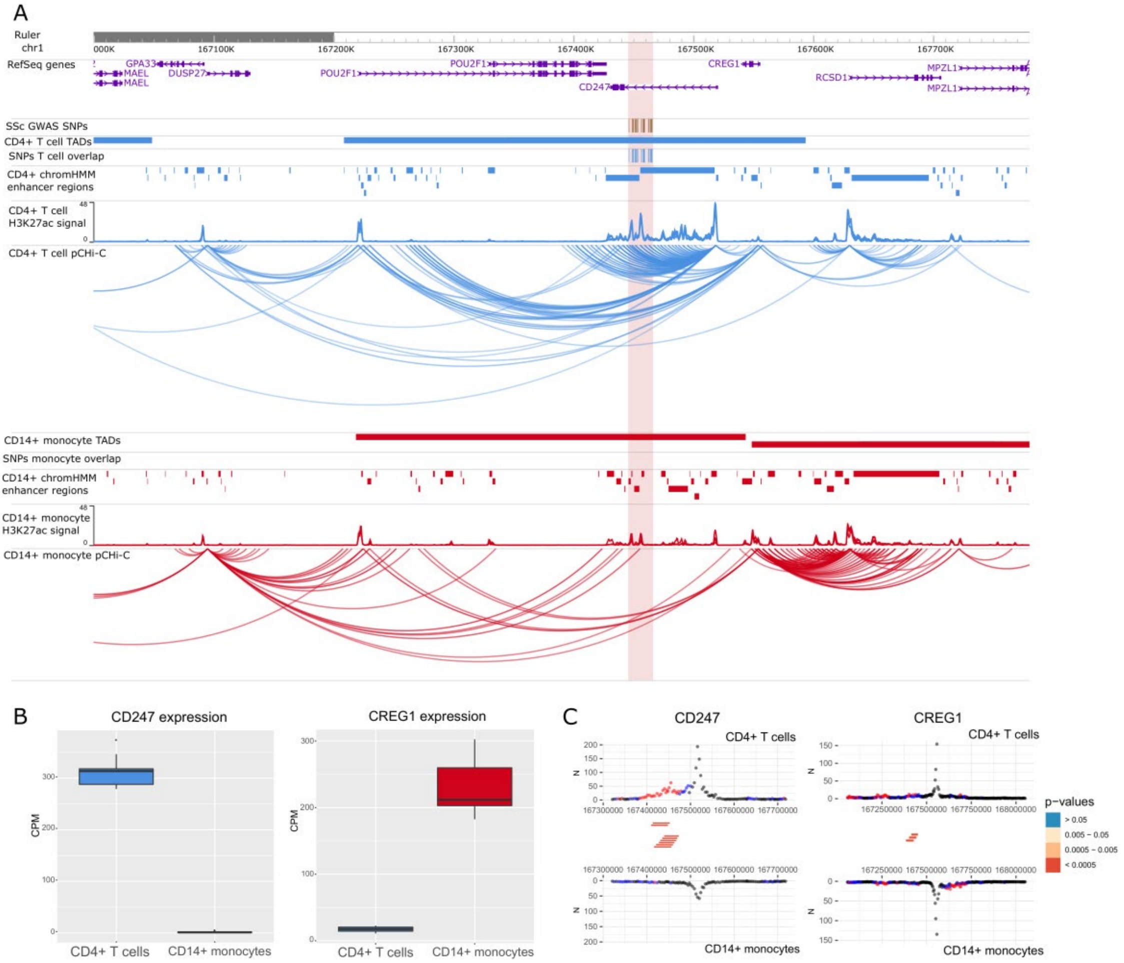


Figure 5.4. Promoter capture Hi-C (pCHi-C) interactions and gene expression in the *CD247* locus. (**A**) Genomic coordinates (GRCh38) are shown at the top of the panel. The tracks include RefSeq genes (NCBI), systemic sclerosis (SSc)-associated SNPs from López-Isac *et al* (106) and their proxies (r^2>0.8), topologically associating domains (TADs) (shown as bars), SNPs overlapping promoter interacting regions (PIRs) and enhancer regions, H3K27ac signal, enhancer regions as defined by chromHMM, and pCHi-C significant interactions (CHiCAGO score > 5) (shown as arcs) in CD4$^+$ T cells (blue) and CD14$^+$ monocytes (red). The highlighted region in red includes all the SSc-associated SNPs LD block. (**B**) Gene expression level of *CD247* and *CREG1* from CD4$^+$ T cells and CD14$^+$ monocytes in counts per million (CPM). (**C**) Chicdiff bait profiles were generated for the *CD247* and *CREG1* genes. Plots show the raw read counts versus linear distance from the bait fragment as mirror images for CD4$^+$ T cells and CD14$^+$ monocytes. Other-end interacting fragments are pooled and color-coded by their adjusted weighted *p*-value. Significant differentially interacting regions detected by Chicdiff overlapping SSc-associated SNPs and enhancer regions are depicted as red blocks.

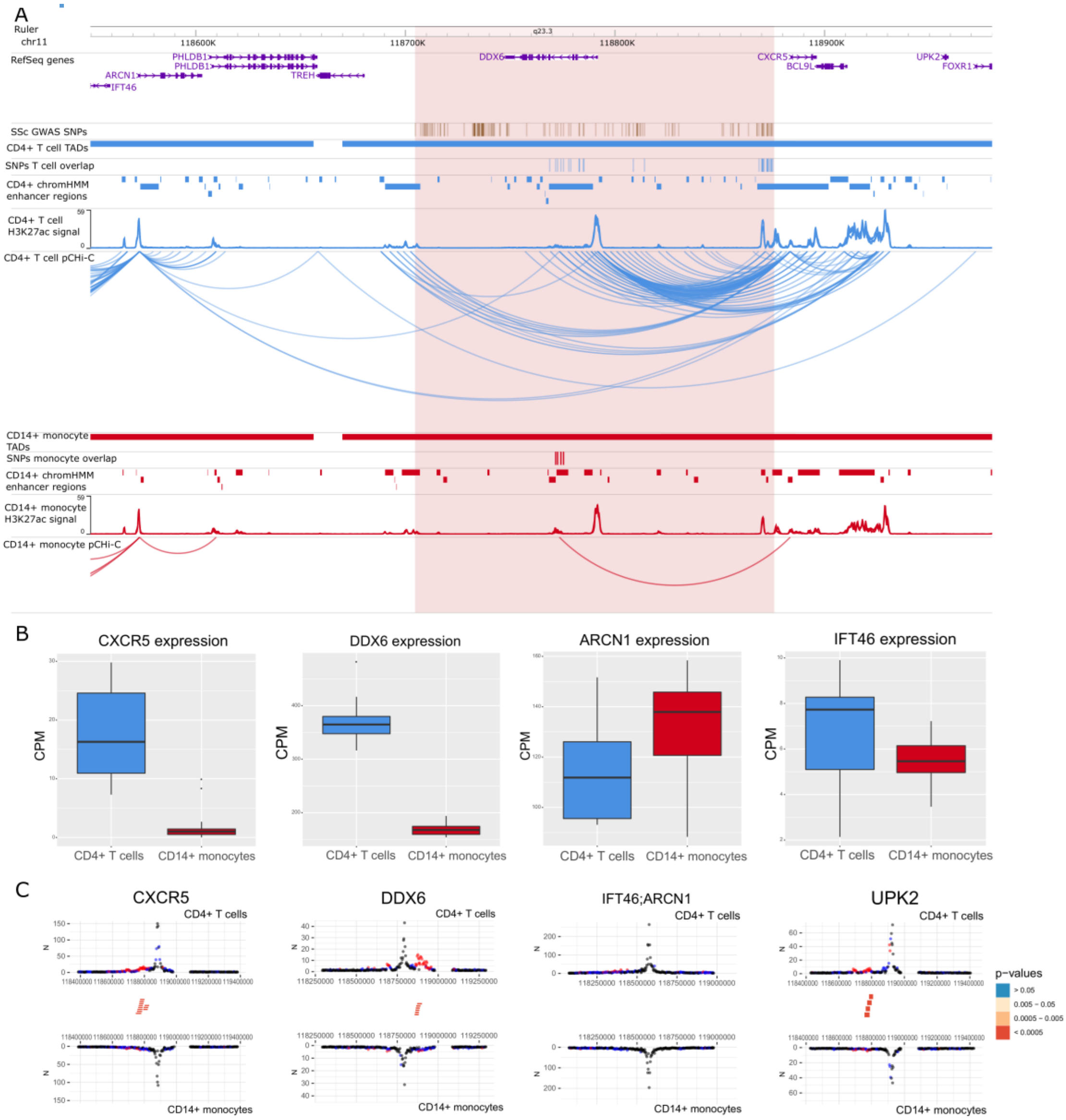


Figure 5.5. Promoter capture Hi-C (pCHi-C) interactions and gene expression in the *DDX6* GWAS locus. (**A**) Genomic coordinates (GRCh38) are shown at the top of the panel. The tracks include RefSeq genes (NCBI), systemic sclerosis (SSc)-associated SNPs from López-Isac *et al* (106) and their proxies (r^2>0.8), topologically associating domains (TADs) (shown as bars), SNPs overlapping promoter interacting regions (PIRs) and enhancer regions, enhancer regions as defined by chromHMM, H3K27ac signal, and pCHi-C significant interactions (CHiCAGO score > 5) (shown as arcs) in CD4$^+$ T cells (blue) and CD14$^+$ monocytes (red). The highlighted region in red includes all the SSc-associated SNPs LD block. (**B**) Gene expression level of *CXCR5*, *DDX6*, *ARCN1*, and *IFT46* from CD4$^+$ T cells and CD14$^+$ monocytes in counts per million (CPM). (**C**) Chicdiff bait profiles were generated for the *CXCR5*, *DDX6*, *IFT46;ARCN1* (shared capture bait), and *UPK2* genes. Plots show the raw read counts versus linear distance from the bait fragment as mirror images for CD4$^+$ T cells and CD14$^+$ monocytes. Other-end interacting fragments are pooled and color-coded by their adjusted weighted *p*-value. Significant differentially interacting regions detected by Chicdiff overlapping SSc-associated SNPs and enhancer regions are depicted as red blocks.

5.2.4. Drug repurposing in SSc

From the 46 genes that present PIRs overlapping significant SSc-associated SNPs and enhancer regions, we identified a total of 21 drugs with interest in SSc targeting protein products in strong PPI with 13 of those genes (five of them specific for CD4^{+} T cells interactions) (**Table 5.2**). Fifteen of these drugs correspond to potential drug targets already in use, or at least in completed clinical phase III, in other similar immune-mediated diseases that could be repurposed for SSc treatment, such as metformin or dimethyl fumarate. Apart from new potential drug targets, tocilizumab was one of the drugs highlighted in our analysis, which in fact represents the only biological drug approved by FDA for its use in SSc-associated interstitial lung disease (266). We also identified five drugs which present advanced clinical trials developed in SSc (tofacitinib, nintedanib, bosentan, methylprednisolone and mycophenolic acid).

Table 5.2. Summary of potential targets for drug repurposing in systemic sclerosis based on pCHi-C data.

GWAS locus	pCHi-C interacting genes	Cell type with interactions	Genes in strong PPI	Targeted drug	Disease indication$^{\$}$
CD247	*CREG1*	CD4+ T cells	*TUBB4B*	Colchicine	Osteoarthritis, Advanced fibrosis
FLNB-DNASE1L3-PXK	*RPP14*	CD4+ T cells, CD14+ monocytes	*KEAP1*	Dimethyl Fumarate	Psoriasis, Multiple sclerosis, Disseminated sclerosis
			AGTR1	Candesartan	Type 1 Diabetes
			HSPA8	Forigerimon	Systemic lupus erythematosus
NFKB1	*NFKB1*	CD4+ T cells	*IL12B*	Ustekinumab	Psoriasis, Crohn´s disease, Ulcerative colitis
			IL1R1	Anakinra	Rheumatoid arthritis
			IL23A	Tildrakizumab	Psoriasis
			JAK2	Tofacitinib	**Systemic sclerosis**, Rheumatoid arthritis, Ulcerative colitis, Interstitial lung disease, Takayasu Arteritis
			NR3C1	Methylprednisolone*	Rheumatoid arthritis, Crohn´s disease, Psoriatic arthritis, Ulcerative colitis, Behcet´s syndrom
	UBE2D3	CD4+ T cells, CD14+ monocytes	*KEAP1*	Dimethyl Fumarate	Psoriasis, Multiple sclerosis, Disseminated sclerosis
RAB2A-CHD7	*SDCBP*	CD4+ T cells	*IMPDH1*	Mycophenolic acid*	Systemic lupus erythematosus, Immunosupresion
			TUBB4B	Colchicine	Osteoarthritis, Advanced fibrosis
	CHD7	CD4+ T cells	*PPARG*	Mesalamine	Crohn's disease, Ulcerative colitis

Only related immune-mediated diseases were included. All clinical trials at least in completed phase III.

*These drugs present phase III or lower clinical trials in systemic sclerosis.

Table 5.2. Summary of potential targets for drug repurposing in systemic sclerosis based on pCHi-C data (continuation).

GWAS locus	pCHi-C interacting genes	Cell type with interactions	Genes in strong PPI	Targeted drug	Disease indication[$]
DDX6	*CXCR5*	CD4+ T cells, CD14+ monocytes	*S1PR3*	Fingolimod	Multiple sclerosis, Disseminated sclerosis
CSK	*CSK*	CD4+ T cells, CD14+ monocytes	*FLT4*	Nintedanib	**Systemic sclerosis**, Idiopathic pulmonary fibrosis, Interstitial lung disease
	COX5A	CD4+ T cells, CD14+ monocytes	*NDUFB10*	Metformin	Type 1 Diabetes, Type 2 Diabetes
IKZF3-GSDMB	*IKZF3*	CD4+ T cells	*JAK1*	Baricitinib	Rheumatoid arthritis
			JAK3	Upadacitinib	Rheumatoid arthritis
			IL2RA	Basiliximab	Type 1 Diabetes
	ERBB2	CD4+ T cells, CD14+ monocytes	*IL6R*	Tocilizumab	**Systemic sclerosis**, Rheumatoid arthritis, Juvenile idiopathic arthritis, Giant cell arteritis
			JAK kinases	Tofacitinib	**Systemic sclerosis**, Rheumatoid arthritis, Ulcerative colitis, Interstitial lung disease, Takayasu Arteritis
IL12RB1	*PIK3R2*	CD4+ T cells	*ADRA1B*	Epinephrine	Crohn's disease
			AGTR1	Candesartan	Type 1 Diabetes
			EDNRA	Bosentan	**Systemic sclerosis**, Idiopathic pulmonary fibrosis, Pulmonary arteria hypertension
			JAK1	Baricitinib	Rheumatoid arthritis
			JAK kinases	Tofacitinib	**Systemic sclerosis**, Rheumatoid arthritis, Ulcerative colitis, Interstitial lung disease, Takayasu Arteritis
			PDGFRB	Nintedanib	**Systemic sclerosis**, Idiopathic pulmonary fibrosis, Interstitial lung disease
	RAB3A	CD4+ T cells, CD14+ monocytes	*HSPA8*	Forigerimod	Systemic lupus erythematosus

[$] Only related immune-mediated diseases were included. All clinical trials at least in completed phase III.

*These drugs present phase III or lower clinical trials in systemic sclerosis.

5.3. Discussion

Our investigation integrates four dimensions for the study of SSc genetics; GWAS, chromatin conformation, gene expression, and cell-specificity. In this regard, our findings are complementary to previously published data and stress the importance of cell type in the functional interpretation of GWAS associations. Through the first pCHi-C analysis in SSc, we identified new target genes, and confirmed others, for SSc-associated loci in two of the main cell types associated with the disease, $CD4^+$ T cells and $CD14^+$ monocytes. Previous studies have acknowledged the potential of chromosome conformation capture to infer potential candidate genes through the analysis of Hi-ChIP data in $CD4^+$ T cells (106,327). Our study provides the next step: comparative studies in two disease relevant cell types and supports future comparisons in multiple cell types by creating the first pCHi-C dataset on T cells and monocytes in SSc.

One of the new candidate genes observed in pCHi-C data corresponds to CXC chemokine receptor type 5 (*CXCR5*) within the *DDX6* locus (**Figure 5.5**). CXCR5 plays an important role in the differentiation of follicular helper T (Tfh) cells, and is highly expressed in $CD4^+$ and $CD8^+$ T cells (328). In addition, a recently published study observed that Tfh cells ($CD4^+CXCR5^+PD\text{-}1^+$) are increased in SSc, and correlate with the severity of the disease (329). In line with the above, interactions with the promoter of this gene were identified specifically in $CD4^+$ T cells in our study, and transcript levels showed an upregulation in this cell type. Furthermore, *CXCR5* has been associated through GWAS with other similar immune-mediated diseases, such as RA or inflammatory bowel disease (184,330). Thus, *CXCR5* represents a good candidate gene contributing to SSc pathology, with a particular interest in $CD4^+$ T cells. Another interesting example is found in the *RAB2A-CHD7* locus, a recently discovered locus associated with SSc (106). Within this region, we observed significant interactions between SSc-associated SNPs and the closest gene, *CHD7,* in CD4$^+$ T cells. *CHD7* is a chromatin remodeler that has been associated with lymphocyte (and other immune-related cells) counts in blood through GWAS (331), and has been previously proposed as a probable candidate gene in SSc (106). Regarding the *IL12A* locus, we found long-range interactions between SSc-associated SNPs by GWAS and the promoter of *SMC4* in CD14$^+$ monocytes. *SMC* family genes play a central role in organizing and compacting chromosomes. In this line, a recent study showed that *SMC4* promotes an inflammatory innate immune response, which is directly associated with monocyte activity, through enhancing NEMO transcription, an essential modulator of NF-κB (332). Although *IL12A* has been traditionally set as the most probable candidate gene for this association, we did not observe any interaction between SSc-associated SNPs and the promoter of this gene. Here, it is important to note the increased difficulty to identify significant short-range interactions (< 1 Mb) as background read count levels are dependent on the distance between fragments (317). This phenomenon represents a limitation in this kind of studies, as most of the GWAS SNPs are classically related with the closest gene, being these SNPs located within the gene itself in some cases. It is worth mentioning that we did not identify any significant interaction overlapping SSc GWAS SNPs in two of the loci most classically associated with SSc susceptibility, as are *IRF5-TNPO3* or *IL12RB2.* Nevertheless, most of the associated SNPs are located within the gene itself in both cases, which means that they could directly affect gene transcription. In this sense, new high resolution Hi-C methods should help overcome the limitation of detecting very short range interactions (333). This methodology is not optimized to be used at the genome-wide level, but could be used as a fine mapping to study individual loci in more detail.

Regarding previously confirmed causal genes associated with SSc, we described interactions between the *IRF8* promoter and SSc-associated variants that were only present in CD14^{+} monocytes (**Figure 5.2**), corresponding with an upregulated expression of this gene in CD14^{+} monocytes as compared with CD4^{+} T cells. This transcription factor plays an important role in differentiation and regulation of monocytes and macrophages (334). Furthermore, variants in *IRF8* have been associated with monocyte counts across different populations (335). In this regard, a recently published study suggests that a downregulation of *IRF8* in monocytes and macrophages of SSc patients may affect the fibrotic phenotype of the disease (336). In addition, another recent study demonstrated that the deletion of an enhancer region corresponding with our SSc GWAS locus in mice model decreased *Irf8* expression, resulting in an overproduction of inflammatory Ly6c^{+} monocytes (337). Thus, our results confirm the association of this candidate gene with SSc through physical chromatin interactions particularly in CD14^{+} monocytes. In the same line, *CD247* and *STAT4* have been described in previous GWAS as main candidate genes associated with SSc (91,106), in this case interactions were exclusively found in CD4^{+} T cells (**Figures 5.3-4**). These findings are in line with literature, as both genes play an important role particularly in T cell signaling and differentiation (338,339). Thus, our results highlight the importance of associating GWAS signals with the specific cell types in which interactions are found, acting as a lead starting point for follow-up functional studies that can relate these signals with the pathogenesis of the disease.

Another aim of this study was to the identification of differences in the interactome between SSc patients and healthy controls. In this regard, we identified 4,858 differential interactions between SSc patients and healthy controls in CD4^{+} T cells. However, these signals appeared to be unreliable. Unfortunately, we did not identify any significant differential interaction in CD14^{+} monocytes. Our results reveal that 3D chromatin structure is largely preserved between SSc patients and healthy controls at least in CD4^{+} T cells and CD14$^+$ monocytes, which make them difficult to interpret due to lack of statistical power and possible bias. Nevertheless, we observed a significant positive correlation in differential interaction $\log_2$FC values between both cell types, which could indicate that the lack of significant differential interactions in CD14$^+$ monocytes could be due to a low statistical power, and that CD4$^+$ T cells differences may indeed represent a true signal. So far, to our knowledge, there is only a published study in which authors attempt to observe differences at the interaction level between patients and healthy controls in complex disorders, concretely in CD4$^+$ T cells from juvenile idiopathic arthritis patients (340). However, authors described almost no differences at the interactomic level, which supports our hypothesis and underlines the difficulty to describe these subtle differences with current technology and resources. Interestingly, it has been shown that subtle differences in chromatin interactions may be correlated with large functional effects on gene expression (341). Thus, larger studies involving larger sample sizes would be of great interest to uncover the potential importance of these differences in understanding the implication of different cell types in disease pathology.

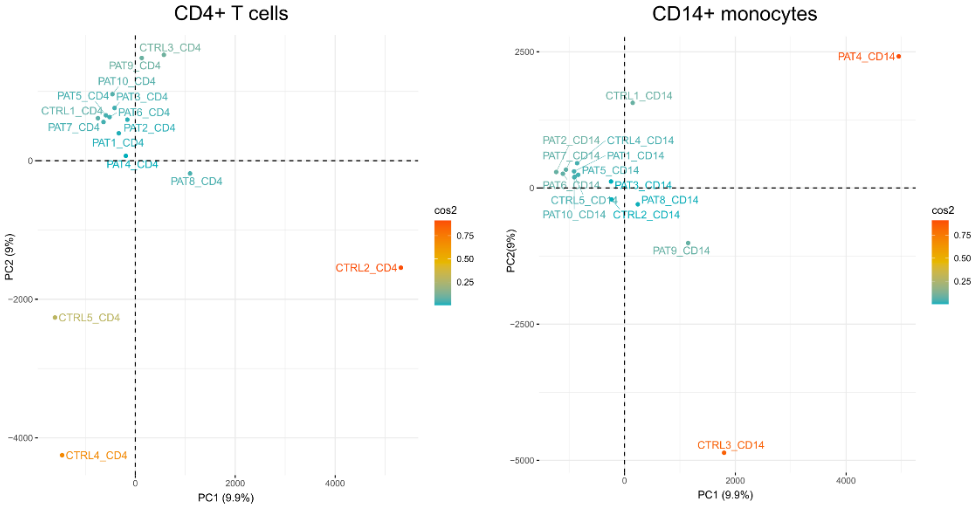
Furthermore, we wanted to describe general differences between cell types at the interaction and expression level, without taking disease status into account. We observed that overexpressed genes in a specific cell type correlated with an increased number of interactions, and that those genes were enriched in specific pathways related with T cells and monocytes signaling, activation, and differentiation. These results demonstrate that interactions are directly related with the expression of important genes implicated in cell type specific pathways. In this regard, a recently published study observed that disease-associated genes tend to be connected by cell-type specific interactions (342). Thus, our data presented here will aid future studies to identify cell types enriched with interactions overlapping GWAS loci.

Finally, it should be noted the relevance of this kind of studies to effectively point to potential drug targets. In this sense, five (23.8%) of the total of 21 drugs of interest identified in our study already present advanced clinical trials developed in SSc, and another one, tocilizumab, represents the only biological drug approved for its use in SSc. These results support the potential for repurposing of the rest of drugs underlined in our study. Among them, fingolimid could represent a good example. This drug promotes down-regulation of sphingosine 1-phosphate (S1P) through S1P receptor modulation, blocking the capacity of lymphocytes to egress from lymph nodes and thus, reducing autoaggressive lymphocyte infiltration (343). The protein encoded by *S1PR3* presents a strong PPI with CXCR5, for which we described strong interactions between *CXCR5* promoter and SSc-associated GWAS SNPs overlapping enhancer regions in the *DDX6* locus, specially in CD4$^+$ T cells (**Figure 5.5**). In addition, we observed a significantly higher *CXCR5* expression in CD4$^+$ T cells as compared to CD14$^+$ monocytes. As previously mentioned, our results suggest that SSc-associated SNPs within the *DDX6* locus could influence susceptibility to the disease through T cell activation and dissemination, which is in fact the mechanism of action targeted by fingolimid, thus, pointing to *CXCR5* as the most likely causal gene for the *DDX6* locus in SSc*,* which could constitute a potential drug target through its PPI with S1PR3.

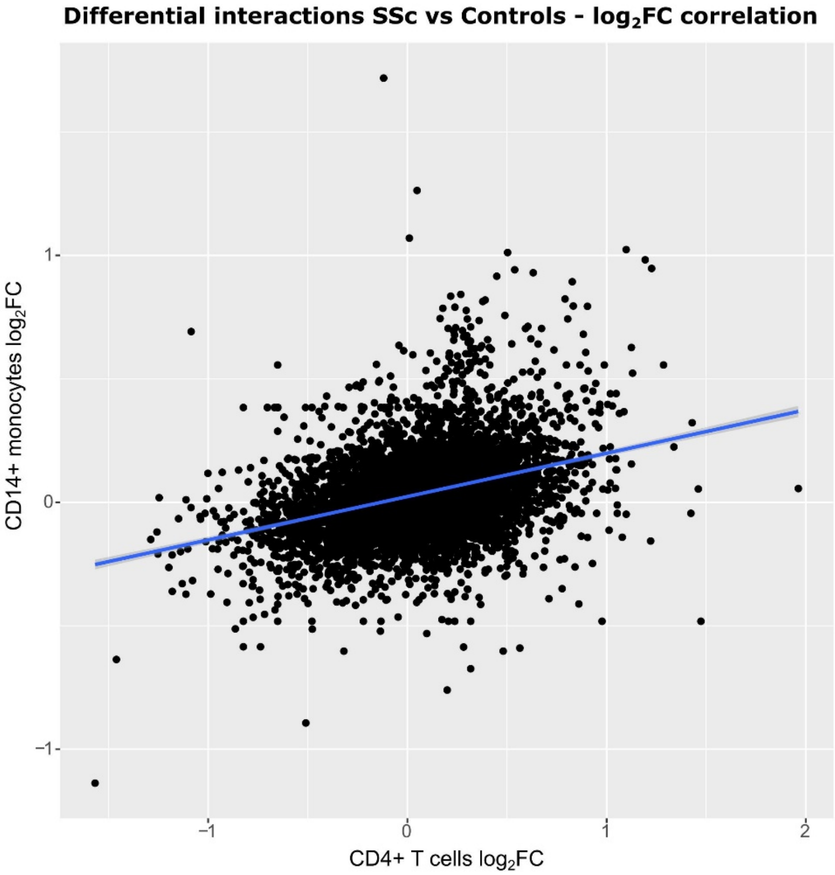
5.4. Supplementary Data

5.4.1. Supplementary Figures

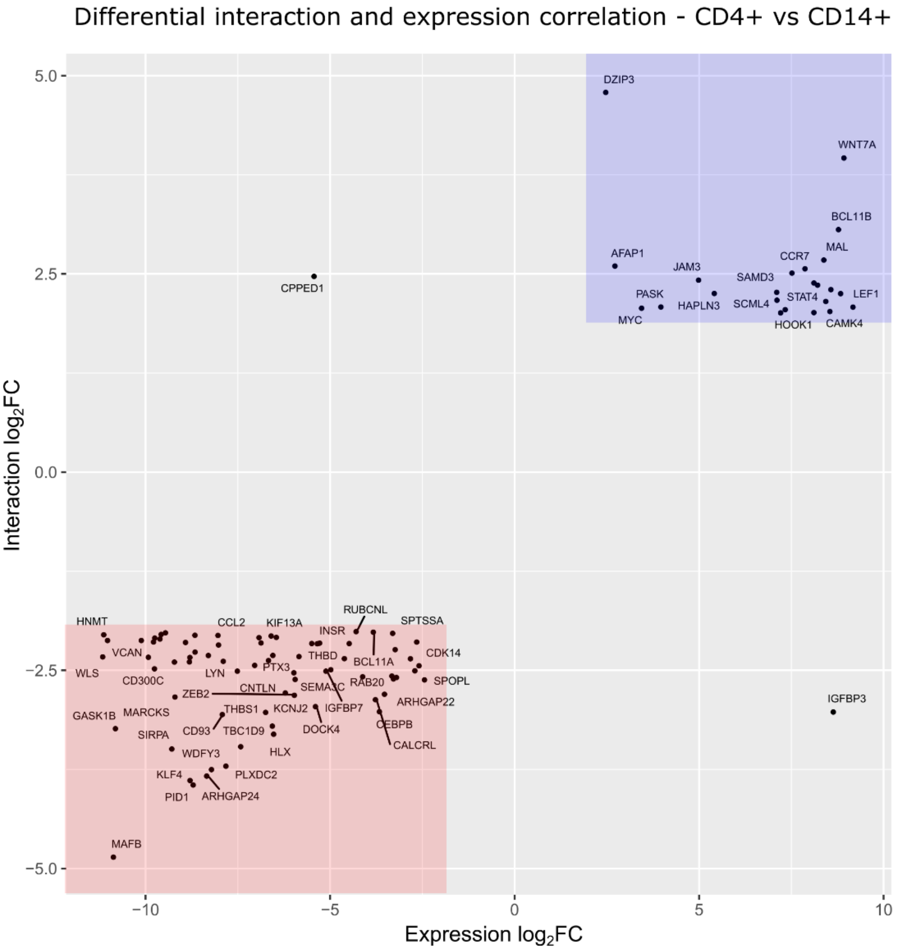
Supplementary Figure S5.1. Representation of the first two principal components (PC) from pCHi-C data for each sample in CD4$^+$ T cells and CD14$^+$ monocytes. The percentage of explained variance of each PC is written in brackets. The cos2 gradient represents the quality of representation (in percentage) of each sample for these two specific PCs.



Supplementary Figure S5.2. Correlation of $\log_2$FC values from systemic sclerosis patients vs healthy control differential interactions between CD4+ T cells and CD14+ monocytes. Each gene is represented by a black dot. The regression line is represented in blue.



Supplementary Figure S5.3. Distribution of the 97 differentially expressed genes (|log2FC| > 2) overlapping differentially interacting genes (|log2FC| > 2) in CD4+ T cells vs CD14+ monocytes comparison. The blue rectangle represents the region including genes which are overexpressed (log2FC > 2) and present more interactions (log2FC > 2) in CD4+ T cells as compared with CD14+ monocytes. The red rectangle represents the region including genes which are overexpressed (log2FC < -2) and present more interactions (log2FC < -2) in CD14+ monocytes as compared with CD4+ T cells.



5.4.2. Supplementary Tables

Supplementary Table S5.1. Clinical and demographic data of systemic sclerosis patients and healthy controls.

Identifier	Disease status	Disease subtype	Age*	Sex
Patient 1	SSc patient	Limited	74	Female
Patient 2	SSc patient	Limited	59	Female
Patient 3	SSc patient	Limited	61	Female
Patient 4	SSc patient	Limited	60	Female
Patient 5	SSc patient	Diffuse	48	Female
Patient 6	SSc patient	Diffuse	36	Female
Patient 7	SSc patient	Diffuse	64	Female
Patient 8	SSc patient	Diffuse	54	Female
Patient 9	SSc patient	Diffuse	64	Female
Patient 10	SSc patient	Limited	64	Female
Control 1	Healthy control	-	26	Male
Control 2	Healthy control	-	50	Female
Control 3	Healthy control	-	50	Female
Control 4	Healthy control	-	60	Male
Control 5	Healthy control	-	56	Female

*Age of individuals when blood extraction was performed

Supplementary Table S5.2. Number of reads after HiCUP allignment and filtering of pCHi-C data.

Identifiers	Cell type	Total read pairs	Paired align	Valid rate (%)	Valid ditags	Unique ditags	Ditags duplication rate (%)	On-target ditags	On-target rate (%)
Control 1	CD4+ T cells	512177179	289850386	82,72	239601506	82654878	65,5	11785577	14,25
	CD14+ monocytes	598399541	407069986	75,44	306989949	58555629	80,93	33553452	57,3
Control 2	CD4+ T cells	559118834	348499840	86,3	300595506	143426522	52,29	39474179	27,52
	CD14+ monocytes	506954077	309516812	86,26	266869550	88746751	66,75	30752664	34,65
Control 3	CD4+ T cells	463266373	282139234	84,26	237678643	95688469	59	31080028	32,48
	CD14+ monocytes	483525548	285998823	86,95	248508586	123620799	50,25	36578861	29,59
Control 4	CD4+ T cells	486749896	316379163	75,72	239415518	44815856	81,28	33944346	75,74
	CD14+ monocytes	545086196	303120064	85,11	257739933	86916088	66,28	27013886	31,08
Control 5	CD4+ T cells	505311506	348101258	71,63	249226756	32832588	86,83	27805005	84,68
	CD14+ monocytes	450094153	251750604	85,84	215892785	81679085	62,17	24476551	29,97
Patient 1	CD4+ T cells	596899626	334134066	85,61	285896447	76754039	73,15	28889836	37,64
	CD14+ monocytes	434717431	248291417	84,83	210506596	82111642	60,99	27694919	33,73
Patient 2	CD4+ T cells	469763189	260633656	85,46	222697560	69414218	68,83	30116005	43,38
	CD14+ monocytes	460010412	253459164	84,73	212830646	46396538	78,2	20670637	44,55
Patient 3	CD4+ T cells	484619244	266203618	84,23	223665186	51908654	76,79	25455852	49,04
	CD14+ monocytes	557578828	306711062	83,88	257078113	73555230	76,39	30411030	41,34

Supplementary Table S5.2. Number of reads after HiCUP allignment and filtering of pCHi-C data (continuation).

Identifiers	Cell type	Total read pairs	Paired align	Valid rate (%)	Valid ditags	Unique ditags	Ditags duplication rate (%)	On-target ditags	On-target rate (%)
Patient 4	CD4+ T cells	469382219	264830498	84,3	223370947	51140806	77,1	30838017	60,3
	CD14+ monocytes	457312653	256003157	86,11	220493190	81500776	63,04	40125407	49,23
Patient 5	CD4+ T cells	487944894	264223927	85,92	227010264	56124422	75,28	22843882	40,7
	CD14+ monocytes	453928751	253398622	85,77	217236914	66179965	69,54	24938066	37,68
Patient 6	CD4+ T cells	448621192	238163644	85,53	203602872	58129875	71,45	24449761	42,06
	CD14+ monocytes	505600799	266272198	85,16	226569823	52073905	77,02	18910095	36,31
Patient 7	CD4+ T cells	534934259	300936296	82,52	248289838	38360709	84,55	21885688	57,05
	CD14+ monocytes	478697222	277297863	84,1	233041721	33566865	85,6	19188795	57,17
Patient 8	CD4+ T cells	556928710	326202981	86,93	283705423	76151835	73,16	36277850	47,64
	CD14+ monocytes	503474455	287223031	86,81	249346399	113587623	54,45	33538229	29,53
Patient 9	CD4+ T cells	482366451	271674962	86,94	236198200	59733249	74,71	29862898	49,99
	CD14+ monocytes	484859062	271899096	88,52	240703905	79402715	67,01	35202959	44,33
Patient 10	CD4+ T cells	430784477	234590464	85,28	200541235	40835308	77,5	21661328	53,05
	CD14+ monocytes	482295416	266056173	82,98	220458961	39118111	81,6	21557477	55,11

Supplementary Table S5.3. Number of reads from RNA-seq data.

Identifiers	Cell type	Read pairs	Unique reads	Duplication rate (%)
Control 2	CD4+ T cells	20,6	2,1	87,99
	CD14+ monocytes	25,15	17,1	39,76
Control 3	CD4+ T cells	29,52	4,87	80,44
	CD14+ monocytes	24,97	18,26	37,63
Control 4	CD4+ T cells	33,67	3,38	87,36
	CD14+ monocytes	3,18	2,6	10,71
Control 5	CD4+ T cells	37,23	3,51	88,00
	CD14+ monocytes	43,37	21,44	48,95
Patient 1	CD4+ T cells	40,12	5,59	82,93
	CD14+ monocytes	37,03	16,59	52,37
Patient 2	CD4+ T cells	33,18	2,78	89,23
	CD14+ monocytes	35,17	2,22	91,36
Patient 3	CD4+ T cells	28,28	2,52	88,67
	CD14+ monocytes	36,81	8,04	74,23
Patient 4	CD4+ T cells	57,28	9,84	78,77
	CD14+ monocytes	31,84	3,29	87,05
Patient 5	CD4+ T cells	35,37	15,24	53,80
	CD14+ monocytes	37,53	16,6	52,52
Patient 6	CD4+ T cells	38,79	5,2	83,42
	CD14+ monocytes	37,22	9,85	68,80
Patient 7	CD4+ T cells	36,38	3,05	89,01
	CD14+ monocytes	36,23	16,24	53,17
Patient 8	CD4+ T cells	20,04	0,85	93,60
	CD14+ monocytes	35,35	8,98	70,72
Patient 9	CD4+ T cells	27,97	1,27	93,24
	CD14+ monocytes	32,66	3,57	86,49
Patient 10	CD4+ T cells	29,5	2,32	89,61
	CD14+ monocytes	32,37	17,42	46,70

Supplementary Table S5.4. Number of SSc GWAS SNPs and significant interactions in each locus.

Locus	Chr	Start (bp)	End (bp)	N SSc GWAS SNPs	N SSc GWAS SNPs + enhancer overlap		N significant interactions (SSc GWAS SNPs + Enhancer overlap)	
					CD4+	CD14+	CD4+	CD14+
1	1	67326053	67448804	27	10	1	0	0
2	1	167445635	167465040	20	20	11	28	0
3	1	173337507	173391947	99	3	2	0	0
4	2	190642047	190698201	28	18	5	31	4
5	2	191035723	191108308	30	21	2	14	0
6	3	58084620	58482701	157	17	37	2	19
7	3	119384733	119546340	28	1	0	1	0
8	3	160002484	160030580	54	14	13	0	13
9	4	960523	990021	12	5	4	6	6
10	4	102477892	102615256	119	97	16	108	10
11	5	151064651	151080486	17	6	8	0	0
12	6	106181815	106339294	59	14	1	0	0
13	7	128933913	129095960	128	26	35	0	0
14	8	11474517	11544554	42	13	13	0	0
15	8	60638547	60664239	11	1	0	3	0
16	11	554659	619789	22	9	6	0	0
17	11	2311894	2363262	80	3	0	3	0
18	11	118704617	118875175	120	31	24	47	4
19	15	74739180	75148328	216	68	54	133	38
20	16	85932852	85979945	46	21	27	0	12
21	17	39747478	39933464	104	22	12	15	1
22	17	75193533	75279345	61	16	9	0	0
23	19	18068862	18093031	25	9	5	7	2
				1505	445	285	398	109

Supplementary Table S5.5. Number of significant pCHi-C interactions and captured promoters identified by group.

Cell type	Group	N of biological replicates	Significant interactions*	N of captured promoters
CD4+ T cells	Total	15	81624	8193
	Patients	10	67573	7356
	Controls	5	89794	9315
CD14+ monocytes	Total	15	74853	7024
	Patients	10	64333	6654
	Controls	5	67041	7908

*Significant interactions are defined as those with CHICAGO score > 5

Supplementary Table S5.6. Differentially expressed genes in SSc vs. controls CD4^{+} T cells.

Gene	log_2FC	log_2CPM	FDR
GPR15	-2,86	5,53	4,57E-05
NUAK2	-3,44	6,33	4,57E-05
LRRN3	-2,53	5,57	2,08E-04
RPS4Y1	-11,63	5,30	2,08E-04
DDX3Y	-11,47	5,13	2,08E-04
KDM5D	-11,10	4,76	2,08E-04
USP9Y	-10,75	4,41	2,14E-04
UTY	-9,26	2,90	6,12E-04
MOSPD2	-1,07	4,48	7,00E-04
GPR55	-1,38	3,69	3,25E-03
PRKY	-8,53	4,06	8,17E-03
PLXNB2	1,93	2,64	1,89E-02
ZNF208	-0,88	4,24	1,92E-02
B3GAT1	4,72	3,94	2,39E-02
C1orf21	4,45	3,55	2,88E-02
TMEM184C	-0,63	4,53	2,88E-02
ITGAX	2,94	2,77	2,88E-02
FGFBP2	5,63	5,29	3,03E-02
HLA-DPB1	1,28	4,91	3,05E-02
IL6R	-0,50	8,21	3,38E-02
CALR	0,43	8,39	3,84E-02
FCRL6	4,51	4,69	3,87E-02

TRPS1	0,60	5,03	3,87E-02
GSE1	0,72	6,18	3,87E-02
ASF1A	-0,48	5,21	3,88E-02
GZMH	5,05	6,00	4,24E-02
RAB31	1,86	2,14	4,33E-02
BEX3	0,92	4,44	6,33E-02
TP53INP2	1,59	3,27	6,46E-02
NBPF15	-0,52	7,13	6,83E-02
ADGRG1	4,32	5,44	6,83E-02
VPS9D1	0,47	5,40	6,83E-02
TP53INP1	0,66	7,26	6,83E-02
ZNF419	-0,79	3,87	6,83E-02
SNTB1	0,87	5,28	6,83E-02
ZEB2	3,43	5,48	6,83E-02
VAV3	1,15	4,52	6,83E-02
C12orf75	1,12	4,68	6,83E-02
FGR	2,93	4,32	6,83E-02
YBX3	1,42	3,77	6,99E-02
ZMIZ1	0,65	6,07	7,00E-02
YIPF4	-0,48	5,84	7,39E-02
FCRL3	1,07	5,99	7,39E-02
FADS2	3,26	3,14	7,39E-02
LGR6	4,52	2,51	7,39E-02
CD36	5,11	0,98	7,39E-02
ADAM28	1,62	2,24	7,39E-02
MS4A1	1,52	3,59	7,39E-02
VCAN	2,25	2,40	7,39E-02
TMEM119	1,85	1,87	8,43E-02
NECTIN2	4,41	0,39	8,43E-02
CDC42BPB	-0,69	4,49	8,83E-02
GNLY	3,78	6,66	8,83E-02
ITGAM	2,40	5,55	8,83E-02
CD74	0,67	8,65	9,17E-02
IL7R	-0,41	11,12	9,17E-02
DUSP6	1,53	2,87	9,23E-02
PIK3AP1	1,72	2,79	9,23E-02
FBLN5	-1,24	4,31	9,23E-02
GFOD1	1,54	3,14	9,23E-02
COL5A3	-1,04	4,24	9,52E-02
GNB4	2,15	1,55	9,52E-02

CPM Counts per million, *FC* fold change, *FDR* false discovery rate.

Supplementary Table S5.7. Differentially expressed genes in SSc vs. controls CD14^{+} T cells.

Genes	log_2FC	log_2CPM	FDR
SEMA6B	-3,14	3,26	8,26E-05
CLEC10A	-1,03	6,08	7,16E-04
STAG3	-1,93	0,13	2,78E-03
RPS4Y1	-12,01	3,77	8,56E-03
KDM5D	-11,29	2,79	9,34E-03
CTTNBP2	-2,29	1,75	9,34E-03
DDX3Y	-9,83	3,60	1,68E-02
CSRNP2	-0,57	4,25	1,68E-02
UTY	-9,74	0,80	1,68E-02
CHAMP1	-0,59	4,15	1,68E-02
TBCC	0,62	4,51	1,68E-02
RHPN1	-1,29	1,18	2,67E-02
USP9Y	-8,53	-0,31	4,17E-02
ACCS	-0,99	5,00	4,17E-02
TIGD2	-0,92	1,94	4,23E-02
ZNF552	-0,75	2,75	4,23E-02
ZNF613	-1,02	2,07	4,23E-02
CYP4F22	2,40	2,85	4,23E-02
FAM118B	-0,56	4,54	4,93E-02
PRKY	-8,03	0,51	5,27E-02
SUOX	-0,47	4,33	5,27E-02
FPR3	-1,37	3,85	5,27E-02
AHRR	-4,28	2,37	5,27E-02
ZNF2	-1,12	1,48	5,27E-02
CMTM6	0,40	8,51	5,27E-02
CAD	-0,54	4,51	5,27E-02
BBS2	-0,48	5,48	5,28E-02
DBR1	-0,67	3,90	5,28E-02
NUAK2	-1,92	5,91	5,82E-02
RAB5IF	0,51	4,85	5,82E-02
SLAMF7	1,31	5,75	5,82E-02
POLH	-0,61	3,24	5,82E-02
MSH6	-0,54	4,64	5,82E-02
LGALS1	0,49	8,01	5,82E-02
MRC1	-1,21	2,43	6,59E-02
GADD45B	0,86	6,27	6,59E-02
GBP2	0,75	8,52	6,59E-02
KLF10	1,12	7,94	6,59E-02
RBM12B	-0,56	4,32	6,59E-02
RNF149	0,48	8,41	6,59E-02
TMEM79	-0,72	3,21	6,59E-02
UBALD2	0,70	6,57	6,59E-02

RSL24D1	0,37	6,76	6,70E-02
APOL1	0,64	5,54	6,70E-02
STX11	0,53	8,61	6,70E-02
ZNF793	-2,26	-0,96	6,70E-02
CLU	2,02	4,12	6,70E-02
ZNF737	-1,11	1,39	6,70E-02
TMEM86A	-0,81	3,19	6,70E-02
BRD8	-0,69	5,22	7,23E-02
ZNF577	-0,64	3,57	7,23E-02
FXR2	-0,55	4,97	7,41E-02
ZC3H13	-0,49	6,43	7,53E-02
SMURF1	0,52	5,51	7,92E-02
FKBP1C	1,31	2,38	7,98E-02
GPR174	1,65	2,08	8,91E-02
MFHAS1	0,95	1,70	8,91E-02
EHHADH	-0,89	1,45	9,44E-02
ZNF160	-0,60	4,76	9,81E-02
RAP1B	0,44	7,44	9,81E-02
CDKN2D	0,77	6,74	9,81E-02
PID1	-0,77	6,57	9,81E-02
ZNF132	-1,47	0,59	9,92E-02

CPM Counts per million, *FC* fold change, *FDR* false discovery rate.

Supplementary Table S5.8. Gene set enrichment analysis of differentially expressed genes in SSc vs. controls CD4$^+$ T cells.

Source	Term name	Term id	Adjusted *p*-value
GO:BP	cell migration	GO:0016477	1,78E-03
GO:BP	positive regulation of immune system process	GO:0002684	5,04E-03
GO:BP	positive regulation of leukocyte activation	GO:0002696	6,96E-03
GO:BP	cell motility	GO:0048870	7,85E-03
GO:BP	localization of cell	GO:0051674	7,85E-03
GO:BP	positive regulation of cell activation	GO:0050867	8,77E-03
GO:BP	leukocyte activation	GO:0045321	1,27E-02
GO:BP	positive regulation of response to stimulus	GO:0048584	1,80E-02
GO:BP	positive regulation of gene expression	GO:0010628	1,82E-02
GO:BP	B cell proliferation	GO:0042100	2,42E-02
GO:BP	locomotion	GO:0040011	3,29E-02
GO:BP	immune effector process	GO:0002252	4,67E-02
GO:BP	cell adhesion	GO:0007155	4,93E-02
GO:CC	side of membrane	GO:0098552	6,91E-05
GO:CC	cell surface	GO:0009986	5,13E-04
GO:CC	external side of plasma membrane	GO:0009897	9,00E-04
GO:CC	endocytic vesicle membrane	GO:0030666	6,00E-03
GO:CC	integral component of lumenal side of endoplasmic reticulum membrane	GO:0071556	1,53E-02
GO:CC	lumenal side of endoplasmic reticulum membrane	GO:0098553	1,53E-02
GO:CC	endocytic vesicle	GO:0030139	1,69E-02
GO:CC	cell periphery	GO:0071944	1,95E-02
GO:CC	lumenal side of membrane	GO:0098576	3,56E-02
KEGG	Hematopoietic cell lineage	KEGG:04640	6,45E-05
WP	Apoptosis-related network due to altered Notch3 in ovarian cancer	WP:WP2864	1,02E-02

Only significant terms with adjusted *p*-value < 0,05 were included.
GO:BP Gene ontology biological process, *GO:CC* Gene ontology cellular component, *KEGG* KEGG pathways, *REAC* Reactome pathways, *WK* WikiPathways.

Supplementary Table S5.9. Gene set enrichment analysis of overexpressed genes in CD4$^+$ T cells.

Source	Term name	Term id	Adjusted *p*-value
GO:MF	T cell receptor binding	GO:0042608	4,46E-04
GO:MF	phosphotransferase activity, alcohol group as acceptor	GO:0016773	8,56E-04
GO:MF	kinase activity	GO:0016301	1,03E-03
GO:MF	beta-catenin binding	GO:0008013	1,80E-03
GO:MF	interleukin-2 receptor activity	GO:0004911	4,65E-03
GO:BP	T cell activation	GO:0042110	1,42E-26
GO:BP	T cell differentiation	GO:0030217	1,89E-23
GO:BP	lymphocyte activation	GO:0046649	2,21E-22
GO:BP	lymphocyte differentiation	GO:0030098	6,06E-22
GO:BP	positive regulation of leukocyte cell-cell adhesion	GO:1903039	1,27E-20
GO:CC	immunological synapse	GO:0001772	3,00E-11
GO:CC	plasma membrane	GO:0005886	1,19E-09
GO:CC	cell surface	GO:0009986	1,00E-07
GO:CC	alpha-beta T cell receptor complex	GO:0042105	3,95E-07
GO:CC	T cell receptor complex	GO:0042101	5,07E-06
KEGG	T cell receptor signaling pathway	KEGG:04660	2,51E-11
KEGG	Th1 and Th2 cell differentiation	KEGG:04658	1,91E-08
KEGG	Th17 cell differentiation	KEGG:04659	3,40E-08
KEGG	PD-L1 expression and PD-1 checkpoint pathway in cancer	KEGG:05235	2,72E-07
KEGG	Primary immunodeficiency	KEGG:05340	2,42E-06
REAC	Generation of second messenger molecules	REAC:R-HSA-202433	1,52E-06
REAC	Translocation of ZAP-70 to Immunological synapse	REAC:R-HSA-202430	4,56E-06
REAC	TCR signaling	REAC:R-HSA-202403	6,40E-06
REAC	Binding of TCF/LEF:CTNNB1 to target gene promoters	REAC:R-HSA-4411364	8,07E-05
REAC	Costimulation by the CD28 family	REAC:R-HSA-388841	8,58E-05
WP	T-Cell antigen Receptor (TCR) Signaling Pathway	WP:WP69	2,43E-10
WP	T-Cell antigen Receptor (TCR) pathway during Staphylococcus aureus infection	WP:WP3863	1,44E-08
WP	T-Cell Receptor and Co-stimulatory Signaling	WP:WP2583	9,93E-08
WP	Cancer immunotherapy by PD-1 blockade	WP:WP4585	1,67E-05

Only first five terms of each source with adjusted *p*-value < 0,05 were included.
GO:MF Gene ontology molecular function, *GO:BP* Gene ontology biological process, *GO:CC* Gene ontology cellular component, *KEGG* KEGG pathways, *REAC* Reactome pathways, *WK* WikiPathways.

Supplementary Table S5.10. Gene set enrichment analysis of overexpressed genes in CD14⁺ monocytes.

Source	Term name	Term id	Adjusted *p*-value
GO:MF	immune receptor activity	GO:0140375	4,93E-11
GO:MF	identical protein binding	GO:0042802	4,91E-10
GO:MF	lipid binding	GO:0008289	1,29E-09
GO:MF	carbohydrate binding	GO:0030246	9,96E-09
GO:MF	pattern recognition receptor activity	GO:0038187	2,46E-08
GO:BP	myeloid leukocyte activation	GO:0002274	4,31E-90
GO:BP	cell activation involved in immune response	GO:0002263	3,06E-79
GO:BP	leukocyte activation involved in immune response	GO:0002366	4,27E-78
GO:BP	leukocyte activation	GO:0045321	3,14E-77
GO:BP	leukocyte degranulation	GO:0043299	7,48E-76
GO:CC	intracellular vesicle	GO:0097708	2,26E-61
GO:CC	cytoplasmic vesicle	GO:0031410	4,62E-61
GO:CC	secretory granule	GO:0030141	6,88E-56
GO:CC	vesicle	GO:0031982	5,66E-51
GO:CC	secretory vesicle	GO:0099503	3,33E-49
KEGG	Tuberculosis	KEGG:05152	6,86E-15
KEGG	Phagosome	KEGG:04145	5,11E-14
KEGG	Leishmaniasis	KEGG:05140	1,49E-13
KEGG	Lysosome	KEGG:04142	1,53E-13
KEGG	Rheumatoid arthritis	KEGG:05323	1,58E-11
REAC	Neutrophil degranulation	REAC:R-HSA-6798695	1,63E-60
REAC	Immune System	REAC:R-HSA-168256	1,67E-52
REAC	Innate Immune System	REAC:R-HSA-168249	2,26E-50
REAC	Toll-like Receptor Cascades	REAC:R-HSA-168898	8,14E-10
REAC	Interleukin-10 signaling	REAC:R-HSA-6783783	5,02E-09
WP	TYROBP Causal Network	WP:WP3945	3,53E-14
WP	Microglia Pathogen Phagocytosis Pathway	WP:WP3937	5,62E-08
WP	IL1 and megakaryocytes in obesity	WP:WP2865	5,75E-08
WP	Vitamin D Receptor Pathway	WP:WP2877	6,52E-06
WP	Human Complement System	WP:WP2806	2,49E-05

Only first five terms of each source with adjusted *p*-value < 0,05 were included.

GO:MF Gene ontology molecular function, *GO:BP* Gene ontology biological process, *GO:CC* Gene ontology cellular component, *KEGG* KEGG pathways, *REAC* Reactome pathways, *WK* WikiPathways.

Supplementary Table S5.11. Gene set enrichment analysis of overexpressed genes overlapping genes with differential interactions in CD4$^+$ T cells.

Source	Term name	Term id	Adjusted *p*-value
GO:BP	T cell differentiation in thymus	GO:0033077	3,65E-03
GO:BP	mononuclear cell differentiation	GO:1903131	1,77E-02
GO:BP	T cell differentiation	GO:0030217	2,26E-02
GO:BP	somatic diversification of T cell receptor genes	GO:0002568	2,34E-02
GO:BP	somatic recombination of T cell receptor gene segments	GO:0002681	2,34E-02
GO:BP	T cell receptor V(D)J recombination	GO:0033153	2,34E-02
REAC	Binding of TCF/LEF:CTNNB1 to target gene promoters	REAC:R-HSA-4411364	1,17E-02
REAC	RUNX3 regulates WNT signaling	REAC:R-HSA-8951430	1,17E-02
WP	Thymic Stromal LymphoPoietin (TSLP) Signaling Pathway	WP:WP2203	3,21E-03
WP	ncRNAs involved in Wnt signaling in hepatocellular carcinoma	WP:WP4336	2,18E-02
WP	LncRNA involvement in canonical Wnt signaling and colorectal cancer	WP:WP4258	3,35E-02
WP	Wnt Signaling Pathway and Pluripotency	WP:WP399	3,65E-02
WP	DNA Damage Response (only ATM dependent)	WP:WP710	4,52E-02
WP	Wnt Signaling	WP:WP428	4,88E-02

Only significant terms with adjusted *p*-value < 0,05 were included.

GO:BP Gene ontology biological process, *KEGG* KEGG pathways, *REAC* Reactome pathways, *WK* WikiPathways.

Supplementary Table S5.12. Gene set enrichment analysis of overexpressed genes overlapping genes with differential interactions in CD14⁺ monocytes.

Source	Term name	Term id	Adjusted *p*-value
GO:MF	opsonin binding	GO:0001846	1,30E-02
GO:MF	complement binding	GO:0001848	3,73E-02
GO:BP	response to lipid	GO:0033993	3,66E-07
GO:BP	leukocyte activation	GO:0045321	5,57E-07
GO:BP	myeloid leukocyte activation	GO:0002274	6,33E-07
GO:BP	cell activation	GO:0001775	9,04E-07
GO:BP	response to oxygen-containing compound	GO:1901700	3,25E-06
GO:CC	endomembrane system	GO:0012505	1,31E-03
GO:CC	cytoplasmic vesicle	GO:0031410	6,31E-03
GO:CC	intracellular vesicle	GO:0097708	6,52E-03
GO:CC	tertiary granule	GO:0070820	6,97E-03
GO:CC	secretory granule	GO:0030141	1,27E-02
KEGG	Malaria	KEGG:05144	1,70E-04
KEGG	Rheumatoid arthritis	KEGG:05323	3,08E-03
KEGG	IL-17 signaling pathway	KEGG:04657	3,82E-03
KEGG	Legionellosis	KEGG:05134	8,58E-03
KEGG	Leishmaniasis	KEGG:05140	2,13E-02
REAC	Neutrophil degranulation	REAC:R-HSA-6798695	1,11E-03
REAC	Regulation of TLR by endogenous ligand	REAC:R-HSA-5686938	2,20E-02
REAC	Immune System	REAC:R-HSA-168256	3,14E-02
REAC	Innate Immune System	REAC:R-HSA-168249	3,49E-02
WP	Platelet-mediated interactions with vascular and circulating cells	WP:WP4462	9,71E-03
WP	LTF danger signal response pathway	WP:WP4478	1,61E-02
WP	Lung fibrosis	WP:WP3624	3,50E-02
WP	TYROBP Causal Network	WP:WP3945	3,50E-02
WP	Spinal Cord Injury	WP:WP2431	3,96E-02

Only first five terms of each source with adjusted *p*-value < 0,05 were included.

GO:MF Gene ontology molecular function, *GO:BP* Gene ontology biological process, *GO:CC* Gene ontology cellular component, *KEGG* KEGG pathways, *REAC* Reactome pathways, *WK* WikiPathways.

Supplementary Table S5.13. Differential expression corresponding to genes with significant interactions overlapping SSc GWAS loci in CD4$^+$ T cells vs CD14$^+$ monocytes.

Chr	GWAS locus	Gene	log$_2$FC	log$_2$CPM	FDR
1	*CD247*	*CD247*	7,49	7,26	4,00E-210
		CREG1	-3,68	6,98	1,32E-325
2	*NAB1*	*MFSD6*	0,29	5,73	2,29E-02
		NEMP2	2,03	3,63	3,16E-61
		HIBCH	0,33	3,46	7,38E-03
		INPP1	-1,88	3,85	1,29E-39
2	*STAT4*	*STAT4*	7,05	6,43	1,00E-304
		NABP1	-0,31	7,08	2,86E-03
3	*FLNB -DNASE1L3-PXK*	*RPP14*	0,13	4,03	2,46E-01
		KCTD6	-0,99	3,14	3,51E-12
3	*POGLUT1-TIMMDC1-CD80- ARHGAP31*	*TMEM39A*	-0,65	4,95	1,03E-11
		POGLUT1	0,41	5,23	6,48E-05
3	*IL12A*	*SMC4*	1,12	5,89	9,27E-34
		IFT80	2,38	3,98	1,35E-114
4	*DGKQ*	*GAK*	-0,34	7,52	1,74E-08
		TMEM175	-0,28	5,65	5,64E-03
		FGFRL1	0,25	4,68	1,50E-01
4	*NFKB1*	*SLC39A8*	1,84	4,38	3,18E-26
		NFKB1	0,04	7,21	5,86E-01
		UBE2D3	-0,50	8,69	3,89E-17
		CISD2	0,30	4,03	1,73E-02
		SLC9B1	NA	NA	NA
		BDH2	1,94	3,43	6,09E-35
8	*RAB2A-CHD7*	*ASPH*	-1,62	5,25	5,49E-37
		SDCBP	-2,92	8,31	1,14E-117
		CHD7	2,58	5,73	1,51E-59
11	*TSPAN32,CD81-AS1*	*TSSC4*	-0,32	5,15	7,88E-05
11	*DDX6*	*CXCR5*	3,21	3,14	1,05E-09
		UPK2	NA	NA	NA
		DDX6	1,14	8,08	2,38E-83
		IFT46	0,25	2,71	1,08E-01
		ARCN1	-0,21	6,96	3,69E-03
15	*CSK*	*CSK*	-0,97	8,39	1,38E-28
		CLK3	-0,18	7,26	4,00E-03
		ULK3	1,50	6,32	4,45E-47
		SCAMP2	-0,38	7,32	8,05E-11
		MPI	1,56	5,30	3,50E-82
		FAM219B	0,38	6,30	1,03E-06
		COX5A	-1,03	5,95	3,84E-46
		C15orf39	-3,64	7,55	2,83E-169
16	*IRF8*	*IRF8*	-4,47	7,28	3,11E-72
17	*IKZF3-GSDMB*	*IKZF3*	5,58	7,12	1,16E-34
		ERBB2	1,82	2,04	1,34E-18
		PSMD3	-0,58	6,47	2,06E-20
19	*IL12RB1*	*PIK3R2*	NA	NA	NA
		RAB3A	0,70	1,56	1,09E-02

Genes with very low expression that were not analyzed are represented as NA.
CPM Counts per million, *FC* fold change, *FDR* false discovery rate.

Supplementary Table S5.14. Differential interaction values corresponding to genes with significant interactions overlapping SSc GWAS loci in CD4$^+$ T cells vs CD14$^+$ monocytes.

			Chicdiff median values per baited promoter		
Chr	**GWAS locus**	**Baited gene promoters**	**log$_2$FC**	**log$_2$FC SE**	**weighted *p*-adjusted**
1	*CD247*	*CD247*	2,29	0,09	2,51E-129
		CREG1	0,54	0,09	1,76E-08
2	*NAB1*	*MFSD6*	0,73	0,08	2,03E-21
		NEMP2	0,26	0,08	2,81E-03
		HIBCH;INPP1	-0,63	0,10	6,49E-11
2	*STAT4*	*STAT4*	2,07	0,11	1,73E-70
		NABP1	0,96	0,10	2,73E-20
3	*FLNB -DNASE1L3-PXK*	*RPP14*	0,52	0,10	4,58E-06
		KCTD6	0,60	0,07	7,08E-26
3	*POGLUT1-TIMMDC1-CD80- ARHGAP31*	*TMEM39A;POGLUT1*	0,46	0,10	1,45E-05
3	*IL12A*	*SMC4;IFT80*	-1,28	0,12	1,27E-23
4	*DGKQ*	*GAK;TMEM175*	-0,16	0,11	1,77E-01
		FGFRL1	-0,19	0,08	1,92E-02
4	*NFKB1*	*SLC39A8*	1,24	0,09	2,45E-40
		NFKB1	1,31	0,11	7,82E-30
		UBE2D3;CISD2	0,85	0,12	7,61E-12
		SLC9B1	0,38	0,13	2,90E-03
		BDH2	-0,09	0,09	2,35E-01
8	*RAB2A-CHD7*	*ASPH*	1,38	0,13	1,26E-25
		SDCBP	0,84	0,17	5,19E-06
		CHD7	0,11	0,07	1,71E-01
11	*TSPAN32,CD81-AS1*	*TSSC4*	-0,38	0,09	8,43E-05
11	*DDX6*	*CXCR5*	0,97	0,09	1,70E-26
		UPK2	0,87	0,13	1,30E-10
		DDX6	1,56	0,12	2,63E-37
		IFT46;ARCN1	0,00	0,10	1,00E+00
15	*CSK*	*CSK*	0,28	0,11	4,33E-02
		CLK3	0,00	0,10	1,00E+00
		ULK3	0,53	0,13	7,81E-05
		SCAMP2	0,41	0,07	1,37E-07
		MPI	0,27	0,08	2,12E-03
		FAM219B	0,42	0,09	1,32E-05
		COX5A	0,74	0,09	3,02E-13
		C15orf39	-1,61	0,13	2,50E-34
16	*IRF8*	*IRF8*	-1,27	0,09	1,19E-39
17	*IKZF3-GSDMB*	*IKZF3*	2,07	0,12	1,98E-67
		ERBB2	0,83	0,11	8,29E-13
		PSMD3	1,89	0,19	3,75E-24
19	*IL12RB1*	*PIK3R2*	0,52	0,16	2,99E-03
		RAB3A	-0,37	0,12	1,69E-03

FC fold change, *SE* standard error.

Supplementary Table S5.15. Gene set enrichment analysis of genes interacting with SSc GWAS loci overlapping enhancer regions in CD4$^+$ T cells.

Source	**Term name**	**Term id**	**Adjusted *p*-value**
GO:MF	protein C-terminus binding	GO:0008022	2,49E-02
KEGG	Pancreatic cancer	KEGG:05212	3,50E-02
KEGG	Epstein-Barr virus infection	KEGG:05169	4,90E-02

Only first five terms of each source with adjusted *p*-value < 0,05 were included.
GO:MF Gene ontology molecular function, *KEGG* KEGG pathways.

Supplementary Table S5.16. Gene set enrichment analysis of genes interacting with SSc GWAS loci overlapping enhancer regions in CD14$^+$ monocytes.

Source	**Term name**	**Term id**	**Adjusted *p*-value**
GO:MF	transmembrane receptor protein tyrosine kinase activity	GO:0004714	7,22E-03
GO:MF	protein tyrosine kinase activity	GO:0004713	1,22E-02
GO:MF	transmembrane receptor protein kinase activity	GO:0019199	1,25E-02
GO:MF	protein kinase activity	GO:0004672	3,16E-02

Only first five terms of each source with adjusted *p*-value < 0,05 were included.
GO:MF Gene ontology molecular function.

GENERAL DISCUSSION

The study of systemic complex ADs, such as SSc, represents a major challenge at the genetic level, as the pathogenesis is lead by the sum of different factors that, at the same time, do not affect an organ or tissue in particular, but the whole organism. Currently, much effort is dedicated to address the study of these diseases through genetic approaches, from the identification of point mutations to the relationship and regulation of different genome regions, and their consequences in different cellular pathways of interest. Through the last decade, formerly misnamed "junk DNA" has become increasingly relevant, as mutations in non-coding regulatory regions can modulate the expression of a great number of genes, acting as major drivers of human diseases, and thus, generating several regulatory changes that can not be approached by reductionist methodologies (344).

The present PhD dissertation provides evidences of the importance of each step concerning genetic studies in complex polygenic diseases, in which the discovery of a new association by genotyping studies represents only the starting point of a large and intrincate path. In this section, we will briefly discuss how the different results described throughout the five chapters complement and complete, as well as future perspectives and critical next steps in the field of SSc genetics.

As previously mentioned, the study of the genetic basis of ADs in non-European populations represents a major challenge nowadays (124). In this regard, it is worth mentioning the results obtained in the GWAS performed on turkish population, approached in chapter 1 of the present thesis. In this study, we identified a suggestive signal corresponding to the *GOT1-NKX2.3* locus, very close to the genome-wide significance threshold. As this locus has been previously associated to CD and ulcerative colitis, we checked if the index

SNP (rs7095491) reached the statistical threshold in the cross-disease meta-GWAS performed in SSc and CD (chapter 2). Interestingly, this SNP showed a suggestive level of association in the meta-analysis of the discovery cohort (*p*-value = 2.03x10^{-6}, OR = 1.15). Nevertheless, this association may be leaded by the CD cohort (*p*-value = 3.59x10^{-7}, OR = 1.25), as compared with the SSc cohort (p-value= 5.10x10^{-2}, OR = 1.07), which present a weaker association. Furthermore, the SNP rs7095491 acts as an eQTL of *NKX2.3* in thyroid tissue (*p*-value = 3.10x10^{-10}) according to the GTEx database (206), which is in fact the most enriched tissue observed in our SSc eQTL analysis (chapter 4), highlighting the potential implication of thyroid tissue in the pathogenesis of the disease. Although replication in independent cohorts is needed in order to determine if this locus represents a true genetic risk factor for, it is of great importance to remark the potential of performing genomic studies in different populations in order to identify new associated loci and to confirm previous signals.

On the other hand, in the cross-disease meta-GWAS including SSc and CD, we identified *ZBTB9-BAK1* associated for the first time with both disorders as a shared risk (chapter 2). The index SNP of this region, rs68191, affected *BAK1* gene exepression, acting as an eQTL in skin (*p*-value = 7.10x10^{-6}) and musculoskeletal tissue (*p*-value = 1.10x10^{-4}) according to the GTEx database (206), both tissues specially affected in SSc patients. Interestingly, our whole blood eQTL study in SSc pointed to *BAK1* (chapter 4) as one of the main eGenes whose expression is affected by SSc-associated SNPs related with apoptosis, highlighting the potential implication of this process in SSc pathogenesis. In this regard, the SNP rs68191 was not directly associated as an eQTL with *BAK1* in our eQTL study, but a set of 5 SNPs which best explained the expression variance, none of them in high LD with rs68191. It is also worth metioning the importance of checking periodically different databases, such as GTEx, as the information is constantly updated and new studies performed in different tissues are included, as an example, we could observe that the SNP rs68191 acts as an eQTL of *BAK1* thanks to this approach.

Regarding the new shared associations across systemic seropositive IMIDs described in chapter 3, 11 of them were new for SSc, including *DGKQ*. Notably, we observed that *DGKQ* expression was deeply influenced by SSc genetics in our eQTL analysis (chapter 4). Furthermore, the association of this locus with SSc was confirmed in a posterior large meta-GWAS performed by López-Isac *et al* (106). Nevertheless, we could not identify physical interactions between this locus and the promoter of *DGKQ* in our pCHi-C analysis, at least in CD4$^+$ T cells and CD14$^+$ monocytes. As most of the SSc-associated SNPs within this locus are intronic, it is probable that regulation of *DGKQ* gene expression by these SNPs is mediated through other mechanisms different to the direct interaction with the promoter region. Interestingly, physical interactions between SSc-associated variants within the *DGKQ* locus and the promoter region of *FGFRL1* were identified in our pCHi-C analysis. This gene encodes a fibroblast growth factor (FGF) receptor implicated in the FGF signaling pathway, which has a crucial role in homeostasis and in regulating differentiation, proliferation and apoptosis of various cell types, including fibroblasts, which play a crucial role in SSc pathogenesis (345). In addition, this signaling pathway is implicated in fibrosis and inflammation of multiple tissues, leading to PF among other complications (345), which is commonly developed by SSc patients (5). In addition, FGF downstream signaling pathway includes the STAT signaling, which has been largely associated with SSc (326). Thus, SNPs associated with SSc in the *DGKQ* locus could be interacting with the *FGFRL1* promoter and, at the same time, be influencing *DGKQ* expression through other mechanisms. This highlights the importance of avoiding the classical association of one locus-one candidate gene, and underlines the complexity and pleiotropy of human gene regulatory variants.

Drug repurposing is one of the common approaches followed on the studies included in the present thesis. The implementation of this strategy among ADs is still difficult because of our limited knowledge about the pathogenesis of these disorders. However, recent genomic studies are being very useful to identify new pathways and targets for this repurposing. Through this strategy, we have identified more than 20 target genes for drugs already indicated in other similar immune-mediated diseases that could be repurposed for SSc treatment. Among these, tofacitinib was identified as a potential drug to be repurposed in SSc in two of our studies (chapter 3 and 5), and, in fact, this drug is now in clinical trials for SSc. Other interesting drug target genes identified in our studies are *ERAP1* and *ERAP2* (chapter 4), which are key regulators of the peptide repertoire displayed by MHC I to circulating T cells and NK cells (346), and can be targeted by aminopeptidases inhibitors. In addition, these genes have recently been pointed out as potential therapeutic targets for ADs (347).

As it has been described throughout the present dissertation, it is through the combination of GWAS data with different functional genomics techniques (such as eQTL analysis or chromosome conformation capture), that associated variants can be linked with causal genes. pCHi-C analysis can partially be used to validate *cis*-eQTL associations. In this respect, it is important to mention that all eQTLs and pCHi-C analyses approached in the present dissertation were performed taking into account only *cis* interactions. However, it has recently been suggested that approximately 70% of heritability in mRNA expression in complex diseases is due to *trans*-eQTLs (348,349). The difficulty to interpret *trans* regulation results is one of the major challenges that have to be approached in future studies, evidencing the complex interactive network existing in the genome. In this regard, as these are not direct physical interactions, a SNP could be influencing the expression levels of certain gene through an unkown intermediate step. It is thus of great importance to trace other changes in the homeostasis of the cell that could connect *trans* regulation through other omic studies, such as proteomics or metabolomics.

In the shorter term, a critical step is to functionally test the interactions identified between SSc-associated SNPs and their target genes through genome editing tools, such as CRISPR/Cas9. Recent studies have demonstrated that this technology can be used to investigate and validate pathological mechanisms of ADs proposed by different genomic approaches. In this line, it is worth mentioning a CRISPR/Cas9 screening study developed in primary human T cells in which the authors identified genes regulating the TCR response after stimulation (350). The combination of CRISPR/Cas9 with other characterization and functional genomics techniques constitutes the perfect toolset to uncover mechanistic insights of potential functional variants. In this regard, a recent study combining CRISPR/Cas9 and luciferase assays with Hi-C and CHi-C data identified the allele rs13239597-A, which is strongly associated with SSc and SLE, as an allele-specific enhancer regulating *IRF5* expression (351). Thus, one of the main next steps that should be approached by our group would be the experimental characterization of the new SSc-associated loci, in order to check if gene expression changes can be detected in those genes interacting with the loci of interest.

Another promising strategy would be applying single-cell technology to different functional genomics studies in SSc. In this line, our group is currently working on a scRNA-seq study in Th17 cells and fibroblasts from SSc patients, two of the most relevant cell types in SSc pathogenesis. Through this approach, we would be able to detect specific cell subpopulations of particular interest in the disease, as well as to characterize their gene expression signature, and to identify dysregulated pathways. Performing different chromosome conformation capture techniques and eQTL analyses in these particular cell types would also help us improving our knowledge of how SSc-associated loci interact and regulate gene expression.

CONCLUSIONS

1. The previously reported association of the HLA region with systemic sclerosis was confirmed in the Iranian and Turkish populations, lead by the *HLA-DRB1*11:04* and *HLA-DPB1*13:01* classical alleles. A suggestive association for two previously SSc-associated loci in Europeans, *IRF5-TNPO3* and *NFKB1*, was also evidenced.

2. *GOT1-NKX2*.3, a genetic risk factor for other immune-mediated diseases, emerged as a potential candidate locus associated with systemic sclerosis susceptibility in the Turkish population.

3. The analysis of the shared genetic component between systemic sclerosis and Crohn's disease identified four novel shared risk loci, *IL12RB2*, *IRF1*, *ZBTB9-BAK1*, and *STAT3,* not previously associated with systemic sclerosis, except for *IL12RB2*.

4. Functional enrichment analysis identified the IL-12/IL-23 signaling as one of the most relevant common pathways between systemic sclerosis and Crohn's disease.

5. The study of the genetic pleiotropy among four systemic seropositive immune-mediated inflammatory diseases identified 26 common loci for at least two conditions, of which *NAB1*, *DGKQ*, *KPNA4-ARL14*, *LIMK1,* and *PRR12* had not been reported before.

6. The pleiotropic variants identified among the four analyzed disorders and their likely target genes are functionally enriched in relevant immune cells, highlighting the type I interferon signaling as the most relevant common pathway.

7. More than half of the 233 eGenes detected in the expression quantitative trait locus (eQTL) analysis were associated with the most important systemic sclerosis hallmarks, highlighting the crucial role of the apoptotic process.

8. Systemic sclerosis specific eQTLs showed enrichment in motifs for transcription factors, which were differentially regulated in disease relevant tissues including skin, blood and lungs.

9. Physical interaction maps revealed cell-type specific interactions between systemic sclerosis-associated loci and previously confirmed causal genes, such as *IRF8* in CD14$^+$ monocytes, and *CD247* and *STAT4* in CD4$^+$ T cells. In addition, interactions between the *DDX6* locus and *CXCR5* gene promoter highlight the potential role of this gene in systemic sclerosis pathogenesis.

10. Our results revealed that 3D chromatin structure is largely preserved between systemic sclerosis patients and healthy controls in CD4$^+$ T cells and CD14$^+$ monocytes.

11. Through the different studies conducted, more than 20 drug target genes already targeted in similar immune-mediated diseases were identified, thus contributing to the potential repositioning of different drugs for its use in systemic sclerosis treatment.

CONCLUSIONES

1. La asociación de la region HLA con esclerosis sistémica identificada en estudios previos, fue confirmada en población iraní y turca, estando dirigida esa asociación por los alelos clásicos *HLA-DRB1*11:04* y *HLA-DPB1*13:01*. También se detectó una asociación a nivel sugestivo de dos loci previamente asociados con esclerosis sistémica en población europea: *IRF5-TNPO3* y *NFKB1.*

2. *GOT1-NKX2*.3, un factor de riesgo genético asociado a otras enfermedades inmunomediadas, representa un locus de susceptibilidad potencial para el desarrollo de esclerosis sistémica en población turca.

3. Mediante el análisis del componente genético común a la esclerosis sistémica y la enfermedad de Crohn, se identificaron cuatro nuevos loci de riesgo compartidos: *IL12RB2*, *IRF1*, *ZBTB9-BAK1*, y *STAT3*, de los cuales únicamente *IL12RB2* se había asociado previamente con la esclerosis sistémica.

4. El análisis de enriquecimiento en rutas moleculares, de acuerdo al componente genético compartido, evidenció la ruta de señalizacion IL-12/IL-23 como una de las principales vías patogénicas comunes a la esclerosis sistémica y la enfermedad de Crohn.

5. El estudio del componente genético compartido entre cuatro enfermedades sistémicas inflamatorias inmunomediadas identificó 26 loci de riesgo comunes a, al menos, dos de estas enfermedades, de los cuales *NAB1*, *DGKQ*, *KPNA4-ARL14*, *LIMK1* y *PRR12* no han sido descritos previamente.

6. Las variantes de susceptibilidad compartidas entre las cuatro enfermedades analizadas y sus posibles genes diana están enriquecidas funcionalmente en células del sistema inmunitario relevantes para estas enfermedades, destacando la señalización del interferón tipo I como una de las vías comunes de mayor relevancia.

7. Más de la mitad de los 233 eGenes detectados en el análisis de locus de caracter cuantitativo de expresión (eQTL) están asociados con los principales rasgos característicos de la esclerosis sistémica, destacando el papel crucal del proceso de apoptosis.

8. Los eQTLs específicos de la esclerosis sistémica presentan un enriquecimiento en motivos de union de factores de transcripción, los cuales están regulados diferencialmente en tejidos implicados en la patología de la enfermedad, incluyendo piel, sangre y pulmones.

9. El estudio de las interacciones físicas del ADN reveló la existencia de interacciones específicas de tipo celular entre loci asociados a la esclerosis sistémica y genes causales previamente confirmados, como es el caso de *IRF8* en monocitos CD14$^+$, y *CD247* y *STAT4* en linfocitos T CD4$^+$. Además, las interacciones detectadas entre el locus *DDX6* y el promotor del gen *CXCR5* indican la potencial relevancia de este gen en la patogénesis de la enfermedad.

10. Nuestros resultados revelan que la estructura tridimensional de la cromatina en linfocitos T CD4$^+$ y monocitos CD14$^+$ está preservada en gran medida entre pacientes con esclerosis sistémica y controles sanos.

11. A través de los diferentes estudios llevados a cabo, se identificaron más de 20 genes con diana farmacológica que, actualmente, están siendo abordados en otras enfermedades inmunomediadas similares, lo que indica que estos fármacos podrían ser potencialmente últiles para el tratamiento de la esclerosis sistémica.

BIBLIOGRAPHY

1. Theofilopoulos AN, Kono DH, Baccala R. The multiple pathways to autoimmunity. Nat Immunol 2017;18(7):716–24.

2. Ngo ST, Steyn FJ, McCombe PA. Gender differences in autoimmune disease. Front Neuroendocrinol. 2014;35(3):347–69.

3. Moroni L, Bianchi I, Lleo A. Geoepidemiology, gender and autoimmune disease. Autoimmun Rev. 2012;11(6–7):A386–92.

4. Wang L, Wang F-S, Gershwin ME. Human autoimmune diseases: a comprehensive update. J Intern Med. 2015;278(4):369–95.

5. Denton CP, Khanna D. Systemic sclerosis. Lancet. 2017;390(10103):1685–99.

6. Katsumoto TR, Whitfield ML, Connolly MK. The Pathogenesis of Systemic Sclerosis. 2011;6:509–37.

7. Morales-Cárdenas A, Pérez-Madrid C, Arias L, Ojeda P, Mahecha MP, Rojas-Villarraga A, et al. Pulmonary involvement in systemic sclerosis. Autoimmun Rev. 2016;15(11):1094–108.

8. Liakouli V, Cipriani P, Marrelli A, Alvaro S, Ruscitti P, Giacomelli R. Angiogenic cytokines and growth factors in systemic sclerosis. Autoimmun Rev. 2011;10(10):590–4.

9. Mostmans Y, Cutolo M, Giddelo C, Decuman S, Melsens K, Declercq H, et al. The role of endothelial cells in the vasculopathy of systemic sclerosis: A systematic review. Autoimmun Rev. 2017;16(8):774–86.

10. Matucci-Cerinic M, Kahaleh B, Wigley FM. Review: Evidence That Systemic Sclerosis Is a Vascular Disease. Arthritis Rheum. 2013;65(8):1953–62.

11. Bhattacharyya S, Wei J, Varga J. Understanding fibrosis in systemic sclerosis: Shifting paradigms, emerging opportunities. Nat Rev Rheumatol. 2012;8(1):42–54.

12. Varga J, Abraham D. Systemic sclerosis: a prototypic multisystem fibrotic disorder. J Clin Invest. 2007;117(3):557–67.

13. Korman B. Evolving insights into the cellular and molecular pathogenesis of fibrosis in systemic sclerosis. Transl Res. 2019;209:77–89.

14. van Caam A, Vonk M, van den Hoogen F, van Lent P, van der Kraan P. Unraveling SSc Pathophysiology; The Myofibroblast. Front Immunol. 2018;0:2452.

15. Lamouille S, Xu J, Derynck R. Molecular mechanisms of epithelial–mesenchymal transition. Nat Rev Mol Cell Biol. 2014;15(3):178–96.

16. Kanno Y. The Role of Fibrinolytic Regulators in Vascular Dysfunction of Systemic Sclerosis. Int J Mol Sci. 2019;20(3):619.

17. Sfikakis PP, McCune BK, Tsokos M, Aroni K, Vayiopoulos G, Tsokos GC. Immunohistological demonstration of transforming growth factor-beta isoforms in the skin of patients with systemic sclerosis. Clin Immunol Immunopathol. 1993;69(2):199–204.

18. Higley H, Persichitte K, Chu S, Waegell W, Vancheeswaran R, Black C. Immunocytochemical Localization and Serologic Detection of Transforming Growth Factor β1. Arthritis Rheum. 1994;37(2):278–88.

19. Lafyatis R. Transforming growth factor β—at the centre of systemic sclerosis. Nat Rev Rheumatol. 2014;10(12):706–19.

20. Higashiyama H, Yoshimoto D, Kaise T, Matsubara S, Fujiwara M, Kikkawa H, et al. Inhibition of activin receptor-like kinase 5 attenuates Bleomycin-induced pulmonary fibrosis. Exp Mol Pathol. 2007;83(1):39–46.

21. Zhang Y, McCormick LL, Gilliam AC. Latency-Associated Peptide Prevents Skin Fibrosis in Murine Sclerodermatous Graft-Versus-Host Disease, a Model for Human Scleroderma. J Invest Dermatol. 2003;121(4):713–9.

22. Rabquer BJ, Koch AE. Angiogenesis and Vasculopathy in Systemic Sclerosis: Evolving Concepts. Curr Rheumatol Reports. 2011;14(1):56–63.

23. Avouac J, Wipff J, Goldman O, Ruiz B, Couraud PO, Chiocchia G, et al.

Angiogenesis in systemic sclerosis: Impaired expression of vascular endothelial growth factor receptor 1 in endothelial progenitor–derived cells under hypoxic conditions. Arthritis Rheum. 2008;58(11):3550–61.

24. Worrell JC, O'Reilly S. Bi-directional communication: Conversations between fibroblasts and immune cells in systemic sclerosis. J Autoimmun. 2020;113:102526.

25. O'Reilly S, Hügle T, van Laar JM. T cells in systemic sclerosis: a reappraisal. Rheumatology. 2012;51(9):1540–9.

26. Pillai S. T and B lymphocytes in fibrosis and systemic sclerosis. Curr Opin Rheumatol. 2019;31(6):576–81.

27. Tan FK, Zhou X, Mayes MD, Gourh P, Guo X, Marcum C, et al. Signatures of differentially regulated interferon gene expression and vasculotrophism in the peripheral blood cells of systemic sclerosis patients. Rheumatology. 2006;45(6):694–702.

28. York MR, Nagai T, Mangini AJ, Lemaire R, Seventer JM van, Lafyatis R. A macrophage marker, siglec-1, is increased on circulating monocytes in patients with systemic sclerosis and induced by type i interferons and toll-like receptor agonists. Arthritis Rheum. 2007;56(3):1010–20.

29. Brkic Z, Bon L van, Cossu M, Helden-Meeuwsen CG van, Vonk MC, Knaapen H, et al. The interferon type I signature is present in systemic sclerosis before overt fibrosis and might contribute to its pathogenesis through high BAFF gene expression and high collagen synthesis. Ann Rheum Dis. 2016;75(8):1567–73.

30. Valmori D, Raffin C, Raimbaud I, Ayyoub M. Human RORγt+ TH17 cells preferentially differentiate from naive FOXP3+Treg in the presence of lineage-specific polarizing factors. Proc Natl Acad Sci. 2010;107(45):19402–7.

31. Parel Y, Aurrand-Lions M, Scheja A, Dayer J-M, Roosnek E, Chizzolini C. Presence of CD4+CD8+ double-positive T cells with very high

interleukin-4 production potential in lesional skin of patients with systemic sclerosis. Arthritis Rheum. 2007;56(10):3459–67.

32. Sakkas LI, Bogdanos DP. Systemic sclerosis: New evidence re-enforces the role of B cells. Autoimmun Rev. 2016;15(2):155–61.

33. Matsushita T, Hasegawa M, Yanaba K, Kodera M, Takehara K, Sato S. Elevated serum BAFF levels in patients with systemic sclerosis: Enhanced BAFF signaling in systemic sclerosis B lymphocytes. Arthritis Rheum. 2006;54(1):192–201.

34. Hachulla E, co-workers on behalf of the E, Clerson P, co-workers on behalf of the E, Airò P, co-workers on behalf of the E, et al. Value of systolic pulmonary arterial pressure as a prognostic factor of death in the systemic sclerosis EUSTAR population. Rheumatology. 2015;54(7):1262–9.

35. 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):581–90.

36. LeRoy EC, Black C, Fleischmajer R, Jablonska S, Krieg T, Medsger Jr T, et al. Scleroderma (systemic sclerosis): classification, subsets and pathogenesis. J Rheumatol. 1988;15(2):202–5.

37. Van Den Hoogen F, Khanna D, Fransen J, Johnson SR, Baron M, Tyndall A, et al. 2013 classification criteria for systemic sclerosis: An American college of rheumatology/European league against rheumatism collaborative initiative. Ann Rheum Dis. 2013;72(11):1747–55.

38. Mehra S, Walker J, Patterson K, Fritzler MJ. Autoantibodies in systemic sclerosis. Autoimmun Rev. 2013 Jan 1;12(3):340–54.

39. Nihtyanova SI, Denton CP. Autoantibodies as predictive tools in systemic sclerosis. Nat Rev Rheumatol 2010 62. 2010;6(2):112–6.

40. Meyer O. How useful are serum autoantibodies in the diagnosis and prognosis of systemic sclerosis? Clin Rheumatol. 1998;17(3):179–80.

41. Gabrielli A, Avvedimento E V., Krieg T. Scleroderma. N Engl J Med. 2009;360(19):1989–2003.

42. Miyawaki S, Asanuma H, Nishiyama S, Yoshinaga Y. Clinical and serological heterogeneity in patients with anticentromere antibodies. J Rheumatol. 2005;32(8).

43. Walker UA, Tyndall A, Czirják L, Denton C, Farge-Bancel D, Kowal-Bielecka O, et al. Clinical risk assessment of organ manifestations in systemic sclerosis: a report from the EULAR Scleroderma Trials And Research group database. Ann Rheum Dis. 2007;66(6):754–63.

44. Terras S, Hartenstein H, Höxtermann S, Gambichler T, Kreuter A. RNA polymerase III autoantibodies may indicate renal and more severe skin involvement in systemic sclerosis. Int J Dermatol. 2016;55(8):882–5.

45. Nguyen B, Assassi S, Arnett FC, Mayes MD. Association of RNA Polymerase III Antibodies with Scleroderma Renal Crisis. J Rheumatol. 2010;37(5):1068–1068.

46. Mahler M, Hudson M, Bentow C, Roup F, Beretta L, Pilar Simeón C, et al. Autoantibodies to stratify systemic sclerosis patients into clinically actionable subsets. Autoimmun Rev. 2020;19(8):102583.

47. Ranque B, Mouthon L. Geoepidemiology of systemic sclerosis. Autoimmun Rev. 2010;9(5):A311–8.

48. Barnes J, Mayes MD. Epidemiology of systemic sclerosis: Incidence, prevalence, survival, risk factors, malignancy, and environmental triggers. Curr Opin Rheumatol. 2012;24(2):165–70.

49. Zhong L, Pope M, Shen Y, Hernandez JJ, Wu L. Prevalence and incidence of systemic sclerosis: A systematic review and meta-analysis. Int J Rheum Dis. 2019;22(12):2096–107.

50. Chifflot H, Fautrel B, Sordet C, Chatelus E, Sibilia J. Incidence and Prevalence of Systemic Sclerosis: A Systematic Literature Review. Semin Arthritis Rheum. 2008;37(4):223–35.

51. JD R. Ethnicity and race and systemic sclerosis: how it affects susceptibility, severity, antibody genetics, and clinical manifestations. Curr Rheumatol Rep. 2003;5(2):160–7.

52. Arnett FC, Howard RF, Tan F, Moulds JM, Bias WB, Durban E, 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):1362–70.

53. Elhai M, Avouac J, Walker UA, Matucci-Cerinic M, Riemekasten G, Airò P, et al. A gender gap in primary and secondary heart dysfunctions in systemic sclerosis: a EUSTAR prospective study. Ann Rheum Dis. 2016;75(1):163–9.

54. Bramwell B. Diffuse Sclerodermia: Its Frequency; Its Occurrence in Stone-Masons; Its Treatment by Fibrolysin—Elevations of Temperature Due to Fibrolysin Injections. Edinb Med J. 1914;12(5):387.

55. Erasmus L. Scleroderma in goldminers on the Witwatersrand with particular reference to pulmonary manifestations. S Afr J Lab Clin Med. 1957;3:209–31.

56. Rodnan GP, Benedek TG, Medsger TA, Cammarata RJ. The association of progressive systemic sclerosis (scleroderma) with coal miners' pneumoconiosis and other forms of silicosis. Ann Intern Med. 1967;66(2):323–34.

57. Miller FW, Alfredsson L, Costenbader KH, Kamen DL, Nelson LM, Norris JM, et al. Epidemiology of environmental exposures and human autoimmune diseases: Findings from a National Institute of Environmental Health Sciences Expert Panel Workshop. J Autoimmun. 2012;39(4):259–71.

58. Maeda M, Nishimura Y, Kumagai N, Hayashi H, Hatayama T, Katoh M, et al. Dysregulation of the immune system caused by silica and asbestos. J Immunotoxicol. 2010;7(4):268–78.

59. Lee S, Hayashi H, Maeda M, Chen Y, Matsuzaki H, Takei-Kumagai N,

et al. Environmental factors producing autoimmune dysregulation – Chronic activation of T cells caused by silica exposure. Immunobiology. 2012;217(7):743–8.

60. Otsuki T, Maeda M, Murakami S, Hayashi H, Miura Y, Kusaka M, et al. Immunological Effects of Silica and Asbestos. Cell Mol Immunol. 2007;4(4):261–8.

61. Marie I, Gehanno J-F. Environmental risk factors of systemic sclerosis. Semin Immunopathol 2015 375. 2015;37(5):463–73.

62. Reinl W. [Scleroderma caused by trichloroethylene effects]. Zentralbl Arbeitsmed. 1957;7:58–60.

63. Kettaneh A, Al Moufti O, Tiev KP, Chayet C, Tolédano C, Fabre B, et al. Occupational exposure to solvents and gender-related risk of systemic sclerosis: a metaanalysis of case-control studies. J Rheumatol. 2007;34(1).

64. Marie I, Menard JF, Duval-Modeste AB, Joly P, Dominique S, Bravard P, et al. Association of occupational exposure with features of systemic sclerosis. J Am Acad Dermatol. 2015;72(3):456–64.

65. Marie I, Gehanno JF, Bubenheim M, Duval-Modeste AB, Joly P, Dominique S, et al. Prospective study to evaluate the association between systemic sclerosis and occupational exposure and review of the literature. Autoimmun Rev. 2014;13(2):151–6.

66. Gold LS, Ward MH, Dosemeci M, Roos AJ De. Systemic autoimmune disease mortality and occupational exposures. Arthritis Rheum. 2007;56(10):3189–201.

67. Noonan CW, Pfau JC, Larson TC, Spence MR. Nested case-control study of autoimmune disease in an asbestos-exposed population. Environ Health Perspect. 2006;114(8):1243–7.

68. Ferro A, Zebedeo CN, Davis C, Ng KW, Pfau JC. Amphibole, but not chrysotile, asbestos induces anti-nuclear autoantibodies and IL-17 in C57BL/6 mice. J Immunotoxicol. 2014;11(3):283–90.

69. Salazar KD, Copeland CB, Luebke RW. Effects of Libby amphibole asbestos exposure on two models of arthritis in the Lewis rat. J Toxicol Environ Health A. 2012;75(6):351–65.

70. Furuzawa-Carballeda J, Vargas-Rojas MI, Cabral AR. Autoimmune inflammation from the Th17 perspective. Autoimmun Rev. 2007;6(3):169–75.

71. Grossman C, Dovrish Z, Shoenfeld Y, Amital H. Do infections facilitate the emergence of systemic sclerosis? Autoimmun Rev. 2011;10(5):244–7.

72. Ferri C, Azzi A, Magro CM. Parvovirus B19 and systemic sclerosis. Br J Dermatol. 2005;152(4):819–20.

73. Ohtsuka T, Yamazaki S. Increased prevalence of human parvovirus B19 DNA in systemic sclerosis skin. Br J Dermatol. 2004;150(6):1091–5.

74. Lunardi C, Bason C, Navone R, Millo E, Damonte G, Corrocher R, et al. Systemic sclerosis immunoglobulin G autoantibodies bind the human cytomegalovirus late protein UL94 and induce apoptosis in human endothelial cells. Nat Med. 2000;6(10):1183–6.

75. Lunardi C, Dolcino M, Peterlana D, Bason C, Navone R, Tamassia N, et al. Antibodies against Human Cytomegalovirus in the Pathogenesis of Systemic Sclerosis: A Gene Array Approach. PLOS Med. 2005;3(1):e2.

76. Farina A, Cirone M, York M, Lenna S, Padilla C, McLaughlin S, et al. Epstein–Barr Virus Infection Induces Aberrant TLR Activation Pathway and Fibroblast–Myofibroblast Conversion in Scleroderma. J Invest Dermatol. 2014;134(4):954–64.

77. Yazawa N, Fujimoto M, Kikuchi K, Kubo M. High seroprevalence of Helicobacter pylori infection in patients with systemic sclerosis: association with esophageal involvement. J Rheumatol. 1998;25:650–3.

78. Radić M, Martinović Kaliterna D, Bonacin D, Morović Vergles J, Radić J. Correlation between Helicobacter pylori infection and systemic sclerosis activity. Rheumatology. 2010;49(9):1784–5.

79. Arora-Singh RK, Assassi S, del Junco DJ, Arnett FC, Perry M, Irfan U, et al. Autoimmune diseases and autoantibodies in the first degree relatives of patients with systemic sclerosis. J Autoimmun. 2010;35(1):52–7.

80. Arnett FC, Cho M, Chatterjee S, Aguilar MB, Reveille JD, Mayes MD. Familial Occurrence Frequencies and Relative Risks for Systemic Sclerosis (Scleroderma) in Three United States Cohorts. Artheritis Rheum. 2001;44(6):1359–62.

81. Feghali-Bostwick C, Medsger TA, Wright TM. 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):1956–63.

82. Assassi S, Arnett FC, Reveille JD, Gourh P, Mayes MD. Clinical, immunologic, and genetic features of familial systemic sclerosis. Arthritis Rheum. 2007;56(6):2031–7.

83. Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating Missing Heritability for Disease from Genome-wide Association Studies. Am J Hum Genet. 2011;88(3):294–305.

84. Bossini-Castillo L, López-Isac E, Martín J. Immunogenetics of systemic sclerosis: Defining heritability, functional variants and shared-autoimmunity pathways. J Autoimmun. 2015;64:53–65.

85. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Nat 2015 5267571. 2015;526(7571):68–74.

86. Broen JCA, Coenen MJH, Radstake TRDJ. Genetics of Systemic Sclerosis: An Update. Curr Rheumatol Reports 2011 141. 2011;14(1):11–21.

87. Raychaudhuri S, Rich SS. Autoimmunity: insights from human genomics. Curr Opin Immunol. 2012;24(5):513–5.

88. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet 2010 117. 2010;11(7):499–511.

89. Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. PLOS Comput Biol. 2012;8(12):e1002822.

90. Zhou X, Lee JE, Arnett FC, Xiong M, Park MY, Yoo YK, et al. HLA–DPB1 and DPB2 are genetic loci for systemic sclerosis: A genome-wide association study in Koreans with replication in North Americans. Arthritis Rheum. 2009;60(12):3807–14.

91. Radstake TRDJ, 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.

92. Dieudé P, Boileau C, Guedj M, Avouac J, Ruiz B, Hachulla E, et al. Independent replication establishes the CD247 gene as a genetic systemic sclerosis susceptibility factor. Ann Rheum Dis. 2011;70(9):1695–6.

93. Allanore Y, Saad M, Dieudé P, Avouac J, Distler JHW, Amouyel P, et al. Genome-Wide Scan Identifies TNIP1, PSORS1C1, and RHOB as Novel Risk Loci for Systemic Sclerosis. PLOS Genet. 2011;7(7):e1002091.

94. Bossini-Castillo L, Martin JE, Broen J, Simeon CP, Beretta L, Gorlova OY, et al. Confirmation of TNIP1 but not RHOB and PSORS1C1 as systemic sclerosis risk factors in a large independent replication study. Ann Rheum Dis. 2013;72(4):602–7.

95. Bossini-Castillo L, Martin J-E, Broen J, Gorlova O, Simeón 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.

96. Martin J-E, 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.

97. López-Isac E, Bossini-Castillo L, Guerra SG, Denton C, Fonseca C,

Assassi S, et al. Identification of IL12RB1 as a novel systemic sclerosis susceptibility locus. Arthritis Rheumatol. 2014;66(12):3521–3.

98. López-Isac E, Campillo-Davo D, Bossini-Castillo L, Guerra SG, Assassi S, Simeón CP, et al. Influence of TYK2 in systemic sclerosis susceptibility: a new locus in the IL-12 pathway. Ann Rheum Dis. 2016;75(8):1521–6.

99. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747–53.

100. Cortes A, Brown MA. Promise and pitfalls of the Immunochip. Arthritis Res Ther. 2011;13(1):1–3.

101. Mayes MD, Bossini-Castillo L, Gorlova O, Martin JE, Zhou X, Chen W V., et al. Immunochip analysis identifies multiple susceptibility loci for systemic sclerosis. Am J Hum Genet. 2014;94(1):47–61.

102. Zochling J, Newell F, Charlesworth JC, Leo P, Stankovich J, Cortes A, et al. An Immunochip-based interrogation of scleroderma susceptibility variants identifies a novel association at DNASE1L3. Arthritis Res Ther. 2014;16(5):1–7.

103. Angiolilli C, Marut W, van der Kroef M, Chouri E, Reedquist KA, Radstake TRDJ. New insights into the genetics and epigenetics of systemic sclerosis. Nat Rev Rheumatol. 2018;14(11):657–73.

104. Okada Y, Eyre S, Suzuki A, Kochi Y, Yamamoto K. Genetics of rheumatoid arthritis: 2018 status. Ann Rheum Dis. 2019;78(4):446–53.

105. Oparina N, Martínez-Bueno M, Alarcón-Riquelme ME. An update on the genetics of systemic lupus erythematosus. Curr Opin Rheumatol. 2019;31(6):659–68.

106. López-Isac E, Acosta-Herrera M, Kerick M, Assassi S, Satpathy AT, Granja J, et al. GWAS for systemic sclerosis identifies multiple risk loci and highlights fibrotic and vasculopathy pathways. Nat Commun. 2019;10(1):1–

14.

107. Acosta-Herrera M, Kerick M, Lopéz-Isac E, Assassi S, Beretta L, Simeón-Aznar CP, et al. Comprehensive analysis of the major histocompatibility complex in systemic sclerosis identifies differential HLA associations by clinical and serological subtypes. Ann Rheum Dis. 2021;80(8):1040–7.

108. de Bakker PIW, Ferreira MAR, Jia X, Neale BM, Raychaudhuri S, Voight BF. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum Mol Genet. 2008;17(R2):R122–8.

109. Cotsapas C, Voight BF, Rossin E, Lage K, Neale BM, Wallace C, et al. Pervasive Sharing of Genetic Effects in Autoimmune Disease. PLOS Genet. 2011;7(8):e1002254.

110. Cotsapas C, Hafler DA. Immune-mediated disease genetics: the shared basis of pathogenesis. Trends Immunol. 2013;34(1):22–6.

111. Martin JE, Assassi S, Diaz-Gallo LM, Broen JC, Simeon CP, Castellvi I, et al. A systemic sclerosis and systemic lupus erythematosus pan-meta-GWAS reveals new shared susceptibility loci. Hum Mol Genet. 2013;22(19):4021–9.

112. López-Isac E, Martín J-E, Assassi S, Simeón CP, Carreira P, Ortego-Centeno N, et al. Brief Report: IRF4 Newly Identified as a Common Susceptibility Locus for Systemic Sclerosis and Rheumatoid Arthritis in a Cross-Disease Meta-Analysis of Genome-Wide Association Studies. Arthritis Rheumatol. 2016;68(9):2338–44.

113. Márquez A, Kerick M, Zhernakova A, Gutierrez-Achury J, Chen W-M, Onengut-Gumuscu S, et al. Meta-analysis of Immunochip data of four autoimmune diseases reveals novel single-disease and cross-phenotype associations. Genome Med. 2018;10(1):1–13.

114. Ellinghaus D, Jostins L, Spain SL, Cortes A, Bethune J, Han B, et al. Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. Nat Genet 2016 485.

2016;48(5):510–8.

115. Li YR, Li J, Zhao SD, Bradfield JP, Mentch FD, Maggadottir SM, et al. Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases. Nat Med 2015 219. 2015;21(9):1018–27.

116. González-Serna D, Ochoa E, López-Isac E, Julià A, Degenhardt F, Ortego-Centeno N, et al. A cross-disease meta-GWAS identifies four new susceptibility loci shared between systemic sclerosis and Crohn's disease. Sci Reports 2020 101. 2020;10(1):1–11.

117. Wang MH, Picco MF. Crohn's Disease: Genetics Update. Gastroenterol Clin North Am. 2017;46(3):449–61.

118. Bettenworth D, Bokemeyer A, Baker M, Mao R, Parker CE, Nguyen T, et al. Assessment of Crohn's disease-associated small bowel strictures and fibrosis on cross-sectional imaging: a systematic review. Gut. 2019;68(6):1115–26.

119. Danese S, Bonovas S, Lopez A, Fiorino G, Sandborn WJ, Rubin DT, et al. Identification of Endpoints for Development of Antifibrosis Drugs for Treatment of Crohn's Disease. Gastroenterology. 2018;155(1):76–87.

120. Kim J, Chun J, Lee C, Han K, Choi S, Lee J, et al. Increased risk of idiopathic pulmonary fibrosis in inflammatory bowel disease: A nationwide study. J Gastroenterol Hepatol. 2020;35(2):249–55.

121. Herzog EL, Mathur A, Tager AM, Feghali-Bostwick C, Schneider F, Varga J. Review: Interstitial Lung Disease Associated With Systemic Sclerosis and Idiopathic Pulmonary Fibrosis: How Similar and Distinct? Arthritis Rheumatol. 2014;66(8):1967–78.

122. Acosta-Herrera M, Kerick M, González-Serna D, Consortium MG, Consortium SG, Wijmenga C, et al. Genome-wide meta-analysis reveals shared new loci in systemic seropositive rheumatic diseases. Ann Rheum Dis. 2019;78(3):311–9.

123. Wallace B, Vummidi D, Khanna D. Management of connective tissue

diseases associated interstitial lung disease: a review of the published literature. Curr Opin Rheumatol. 2016;28(3):236–45.

124. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. Cell. 2019;177(1):26–31.

125. Arnett FC, Gourh P, Shete S, Ahn CW, Honey RE, Agarwal SK, 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):822–7.

126. Rodriguez-Reyna TS, Mercado-Velázquez P, Yu N, Alosco S, Ohashi M, Lebedeva T, et al. HLA Class I and II Blocks Are Associated to Susceptibility, Clinical Subtypes and Autoantibodies in Mexican Systemic Sclerosis (SSc) Patients. PLoS One. 2015;10(5):e0126727.

127. Terao C, Kawaguchi T, Dieude P, Varga J, Kuwana M, Hudson M, et al. Transethnic meta-Analysis identifies GSDMA and PRDM1 as susceptibility genes to systemic sclerosis. Ann Rheum Dis. 2017;76(6):1150–8.

128. González-Serna D, López-Isac E, Yilmaz N, Gharibdoost F, Jamshidi A, Kavosi H, et al. Analysis of the genetic component of systemic sclerosis in Iranian and Turkish populations through a genome-wide association study. Rheumatology. 2019;58(2):289–98.

129. Farh KK-H, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nat 2014 5187539. 2014 Oct 29;518(7539):337–43.

130. Ding J, Frantzeskos A, Orozco G. Functional genomics in autoimmune diseases. Hum Mol Genet. 2020;29(R1):R59–65.

131. González-Serna D, Villanueva-Martin G, Acosta-Herrera M, Márquez A, Martín J. Approaching Shared Pathophysiology in Immune-Mediated Diseases through Functional Genomics. Genes 2020, Vol 11, Page 1482. 2020;11(12):1482.

132. Assassi S, Swindell WR, Wu M, Tan FD, Khanna D, Furst DE, et al. Dissecting the heterogeneity of skin gene expression patterns in systemic sclerosis. Arthritis Rheumatol. 2015;67(11):3016–26.

133. Pendergrass SA, Lemaire R, Francis IP, Mahoney JM, Lafyatis R, Whitfield ML. Intrinsic Gene Expression Subsets of Diffuse Cutaneous Systemic Sclerosis Are Stable in Serial Skin Biopsies. J Invest Dermatol. 2012;132(5):1363–73.

134. Whitfield ML, Finlay DR, Murray JI, Troyanskaya OG, Chi J-T, Pergamenschikov A, et al. Systemic and cell type-specific gene expression patterns in scleroderma skin. Proc Natl Acad Sci. 2003;100(21):12319–24.

135. Christmann RB, Sampaio-Barros P, Stifano G, Borges CL, Carvalho CR de, Kairalla R, et al. Association of Interferon- and Transforming Growth Factor β–Regulated Genes and Macrophage Activation With Systemic Sclerosis–Related Progressive Lung Fibrosis. Arthritis Rheumatol. 2014;66(3):714–25.

136. Hsu E, Shi H, Jordan RM, Lyons-Weiler J, Pilewski JM, Feghali-Bostwick CA. Lung tissues in patients with systemic sclerosis have gene expression patterns unique to pulmonary fibrosis and pulmonary hypertension. Arthritis Rheum. 2011;63(3):783–94.

137. Bellocchi C, Ying J, Goldmuntz EA, Keyes-Elstein L, Varga J, Hinchcliff ME, et al. Large-Scale Characterization of Systemic Sclerosis Serum Protein Profile: Comparison to Peripheral Blood Cell Transcriptome and Correlations With Skin/Lung Fibrosis. Arthritis Rheumatol. 2021;73(4):660–70.

138. Risbano MG, Meadows CA, Coldren CD, Jenkins TJ, Edwards MG, Collier D, et al. Altered Immune Phenotype in Peripheral Blood Cells of Patients with Scleroderma-Associated Pulmonary Hypertension. Clin Transl Sci. 2010;3(5):210–8.

139. Taroni JN, Martyanov V, Huang C-C, Mahoney JM, Hirano I, Shetuni

B, et al. Molecular characterization of systemic sclerosis esophageal pathology identifies inflammatory and proliferative signatures. Arthritis Res Ther. 2015;17(1).

140. Taroni JN, Greene CS, Martyanov V, Wood TA, Christmann RB, Farber HW, et al. A novel multi-network approach reveals tissue-specific cellular modulators of fibrosis in systemic sclerosis. Genome Med 2017 91. 2017;9(1):1–24.

141. Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, et al. Understanding multicellular function and disease with human tissue-specific networks. Nat Genet. 2015;47(6):569–76.

142. Mahoney JM, Taroni J, Martyanov V, Wood TA, Greene CS, Pioli PA, et al. Systems Level Analysis of Systemic Sclerosis Shows a Network of Immune and Profibrotic Pathways Connected with Genetic Polymorphisms. PLoS Comput Biol. 2015;11(1).

143. Umans BD, Battle A, Gilad Y. Where Are the Disease-Associated eQTLs? Trends Genet. 2021;37(2):109–24.

144. Moreno-Moral A, Bagnati M, Koturan S, Ko J-H, Fonseca C, Harmston N, et al. Changes in macrophage transcriptome associate with systemic sclerosis and mediate GSDMA contribution to disease risk. Ann Rheum Dis. 2018;77(4):596–601.

145. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2008 101. 2009;10(1):57–63.

146. Beretta L, Barturen G, Vigone B, Bellocchi C, Hunzelmann N, Langhe E De, et al. Genome-wide whole blood transcriptome profiling in a large European cohort of systemic sclerosis patients. Ann Rheum Dis. 2020;79(9):1218–26.

147. Kerick M, González-Serna D, Carnero-Montoro E, Teruel M, Acosta-Herrera M, Makowska Z, et al. Expression Quantitative Trait Locus Analysis in Systemic Sclerosis Identifies New Candidate Genes Associated With Multiple

Aspects of Disease Pathology. Arthritis Rheumatol. 2021;73(7):1288–300.

148. Papalexi E, Satija R. Single-cell RNA sequencing to explore immune cell heterogeneity. Nat Rev Immunol 2017 181. 2017;18(1):35–45.

149. Tabib T, Huang M, Morse N, Papazoglou A, Behera R, Jia M, et al. Myofibroblast transcriptome indicates SFRP2hi fibroblast progenitors in systemic sclerosis skin. Nat Commun. 2021;12(1):1–13.

150. Valenzi E, Bulik M, Tabib T, Morse C, Sembrat J, Bittar HT, et al. Single-cell analysis reveals fibroblast heterogeneity and myofibroblasts in systemic sclerosis-associated interstitial lung disease. Ann Rheum Dis. 2019;78(10):1379–87.

151. Gaydosik AM, Tabib T, Domsic R, Khanna D, Lafyatis R, Fuschiotti P. Single-cell transcriptome analysis identifies skin-specific T-cell responses in systemic sclerosis. Ann Rheum Dis. 2021;80:1453-1460.

152. Tsou PS, Sawalha AH. Unfolding the pathogenesis of scleroderma through genomics and epigenomics. J Autoimmun. 2017;83:73–94.

153. Altorok N, Tsou P-S, Coit P, Khanna D, Sawalha AH. Genome-wide DNA methylation analysis in dermal fibroblasts from patients with diffuse and limited systemic sclerosis reveals common and subset-specific DNA methylation aberrancies. Ann Rheum Dis. 2015;74(8):1612–20.

154. Wang Q, Xiao Y, Shi Y, Luo Y, Li YP, Zhao M, et al. Overexpression of JMJD3 may contribute to demethylation of H3K27me3 in CD4 + T cells from patients with systemic sclerosis. Clin Immunol. 2015;161(2):396–9.

155. Kroef M van der, Castellucci M, Mokry M, Cossu M, Garonzi M, Bossini-Castillo LM, et al. Histone modifications underlie monocyte dysregulation in patients with systemic sclerosis, underlining the treatment potential of epigenetic targeting. Ann Rheum Dis. 2019;78(4):529–38.

156. Tsou P-S, Campbell P, Amin MA, Coit P, Miller S, Fox DA, et al. Inhibition of EZH2 prevents fibrosis and restores normal angiogenesis in scleroderma. Proc Natl Acad Sci. 2019;116(9):3695–702.

157. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nat 2012 4897414. 2012;489(7414):109–13.

158. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nat 2015 5187539. 2015;518(7539):317–30.

159. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell. 2014;159(7):1665–80.

160. Zeggini E, Gloyn AL, Barton AC, Wain L V. Translational genomics and precision medicine: Moving from the lab to the clinic. Science. 2019;365(6460):1409–13.

161. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science. 2002;295(5558):1306–11.

162. Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326(5950):289–93.

163. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi–C: A comprehensive technique to capture the conformation of genomes. Methods. 2012;58(3):268–76.

164. Schmitt AD, Hu M, Ren B. Genome-wide mapping and analysis of chromosome architecture. Nat Rev Mol Cell Biol. 2016;17(12):743–55.

165. Schoenfelder S, Javierre B-M, Furlan-Magaril M, Wingett SW, Fraser P. Promoter Capture Hi-C: High-resolution, Genome-wide Profiling of Promoter Interactions. Journal Vis Exp. 2018;(136):e57320.

166. Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell. 2016;167(5):1369-1384.e19.

167. Martin P, McGovern A, Orozco G, Duffus K, Yarwood A, Schoenfelder S, et al. Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. Nat Commun 2015 61. 2015;6(1):1–7.

168. Martin P, Ding J, Duffus K, Gaddi VP, McGovern A, Ray-Jones H, et al. Chromatin interactions reveal novel gene targets for drug repositioning in rheumatic diseases. Ann Rheum Dis. 2019;78(8):1127–34.

169. Li T, Ortiz-Fernández L, Andrés-León E, Ciudad L, Javierre BM, López-Isac E, et al. Epigenomics and transcriptomics of systemic sclerosis CD4+ T cells reveal long-range dysregulation of key inflammatory pathways mediated by disease-associated susceptibility loci. Genome Med. 2020;12(1):1–17.

170. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.

171. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48(10):1284–7.

172. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nat Methods. 2012;9(2):179–81.

173. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016;48(10):1279–83.

174. Jia X, Han B, Onengut-Gumuscu S, Chen WM, Concannon PJ, Rich SS, et al. Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens. PLoS One. 2013;8(6).

175. Brown WM, Pierce J, Hilner JE, Perdue LH, Lohman K, Li L, et al. Overview of the MHC fine mapping data. Diabetes, Obes Metab. 2009;11(SUPPL. 1):2–7.

176. Raychaudhuri S, Sandor C, Stahl EA, Freudenberg J, Lee HS, Jia X, 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):291–6.

177. Gorlova O, Martin J-E, Rueda B, Koeleman BPC, 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.

178. Beretta L, Rueda B, Marchini M, Santaniello A, Simeón CP, Fonollosa V, et al. Analysis of Class II human leucocyte antigens in Italian and Spanish systemic sclerosis. Rheumatology. 2012;51(1):52–9.

179. Simeon CP, Fonollosa V, Tolosa C, Palou E, Selva A, Solans R, et al. Association of HLA Class II Genes with Systemic Sclerosis in Spanish Patients. J Rheumatol. 2009;36(12):2733–6.

180. Dai S, Murphy GA, Crawford F, Mack DG, Falta MT, Marrack P, et al. Crystal structure of HLA-DP2 and implications for chronic beryllium disease. Proc Natl Acad Sci. 2010;107(16):7425–30.

181. Ito I, Kawaguchi Y, Kawasaki A, Hasegawa M, Ohashi J, Hikami K, et al. Association of a functional polymorphism in the IRF5 region with systemic sclerosis in a Japanese population. Arthritis Rheum. 2009;60(6):1845–50.

182. Dieudé 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.

183. 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. 2009;358(9):900–9.

184. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature.

2013;506(7488):376–81.

185. Arismendi M, Giraud M, Ruzehaji N, Dieudé P, Koumakis E, Ruiz B, et al. Identification of NF-κB and PLCL2 as new susceptibility genes and highlights on a potential role of IRF8 through interferon signature modulation in systemic sclerosis. Arthritis Res Ther. 2015;17(1):1–11.

186. Sun S-C, Chang J-H, Jin J. Regulation of nuclear factor-κB in autoimmunity. Trends Immunol. 2013;34(6):282–9.

187. Bhattacharyya S, Wang W, Graham LVD, Varga J. A20 suppresses canonical Smad-dependent fibroblast activation: novel function for an endogenous inflammatory modulator. Arthritis Res Ther. 2016;18(1):1–10.

188. Liu JZ, Van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet. 2015;47(9):979–86.

189. de Lange KM, Moutsianas L, Lee JC, Lamb CA, Luo Y, Kennedy NA, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. Nat Genet. 2017;49(2):256–61.

190. Yu W, Hegarty JP, Berg A, Chen X, West G, Kelly AA, et al. NKX2-3 Transcriptional Regulation of Endothelin-1 and VEGF Signaling in Human Intestinal Microvascular Endothelial Cells. PLoS One. 2011;6(5):e20454.

191. Vojkovics D, Kellermayer Z, Kajtár B, Roncador G, Vincze Á, Balogh P. Nkx2-3—A Slippery Slope From Development Through Inflammation Toward Hematopoietic Malignancies. Biomark Insights. 2018;13:1177271918757480

192. Carmona FD, Onat AM, Fernández-Aranguren T, Serrano-Fernández A, Robledo G, Direskeneli H, et al. Analysis of Systemic Sclerosis-associated Genes in a Turkish Population. J Rheumatol. 2016;43(7):1376–9.

193. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, et al. Genome-wide association study of 14,000 cases of seven

common diseases and 3,000 shared controls. Nature. 2007;447(7145):661–78.

194. Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet. 2010;42(12):1118–25.

195. Julià A, Domènech E, Ricart E, Tortosa R, García-Sánchez V, Gisbert JP, et al. A genome-wide association study on a southern European population identifies a new Crohn's disease susceptibility locus at RBX1-EP300. Gut. 2013;62(10):1440–5.

196. Rioux JD, Xavier RJ, Taylor KD, Silverberg MS, Goyette P, Huett A, et al. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat Genet. 2007;39(5):596–604.

197. Lennard-Jones JE. Classification of inflammatory bowel disease in children. Scand J Gastroenterol. 1989;24:16–9.

198. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 2010;26(16):2069–70.

199. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017;45(D1):D896–901.

200. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet. 2006;38(2):209–13.

201. Machiela MJ, Chanock SJ. LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics. 2014;31(21):3555–7.

202. Ward LD, Kellis M. HaploReg: A resource for exploring chromatin

states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 2012;40(D1):930–4.

203. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47(D1):D607–13.

204. Sun L, He C, Nair L, Yeung J, Egwuagu CE. Interleukin 12 (IL-12) family cytokines: Role in immune pathogenesis and treatment of CNS autoimmune disease. Cytokine. 2015 Oct 1;75(2):249–55.

205. Skaug B, Assassi S. Type I interferon dysregulation in Systemic Sclerosis. Cytokine. 2020;132:154635.

206. Ardlie KG, DeLuca DS, Segrè A V., Sullivan TJ, Young TR, Gelfand ET, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science. 2015;348(6235):648–60.

207. Fujiya M, Inaba Y, Musch MW, Hu S, Kohgo Y, Chang EB. Cytokine regulation of OCTN2 expression and activity in small and large intestine. Inflamm Bowel Dis. 2011;17(4):907–16.

208. Gaowa S, Zhou W, Yu L, Zhou X, Liao K, Yang K, et al. Effect of Th17 and Treg axis disorder on outcomes of pulmonary arterial hypertension in connective tissue diseases. Mediators Inflamm. 2014;2014:247372.

209. Park M-J, Moon S-J, Lee E-J, Jung K-A, Kim E-K, Kim D-S, et al. IL-1-IL-17 Signaling Axis Contributes to Fibrosis and Inflammation in Two Different Murine Models of Systemic Sclerosis. Front Immunol. 2018;0:1611.

210. Hueber W, Sands BE, Lewitzky S, Vandemeulebroecke M, Reinisch W, Higgins PDR, et al. Secukinumab, a human anti-IL-17A monoclonal antibody, for moderate to severe Crohn’s disease: unexpected results of a randomised, double-blind placebo-controlled trial. Gut. 2012;61(12):1693–700.

211. Minegishi Y, Saito M, Tsuchiya S, Tsuge I, Takada H, Hara T, et al.

Dominant-negative mutations in the DNA-binding domain of STAT3 cause hyper-IgE syndrome. Nature. 2007;448(7157):1058–62.

212. Koumakis E, Dieudé P, Vouac J, Kahan A, Allanore Y. Familial Autoimmunity in Systemic Sclerosis — Results of a French-based Case-Control Family Study. J Rheumatol. 2012;39(3):532–8.

213. Tseng C-C, Yen J-H, Tsai W-C, Ou T-T, Wu C-C, Sung W-Y, et al. Reduced incidence of Crohn's disease in systemic sclerosis: a nationwide population study. BMC Musculoskelet Disord. 2015;16(1):1–7.

214. Lin PI, Vance JM, Pericak-Vance MA, Martin ER. No Gene Is an Island: The Flip-Flop Phenomenon. Am J Hum Genet. 2007;80(3):531–8.

215. Wang K, Baldassano R, Zhang H, Qu H-Q, Imielinski M, Kugathasan S, et al. Comparative genetic analysis of inflammatory bowel disease and type 1 diabetes implicates multiple loci with opposite effects. Hum Mol Genet. 2010;19(10):2059–67.

216. Parkes M, Cortes A, van Heel DA, Brown MA. Genetic insights into common pathways and complex relationships among immune-mediated diseases. Nat Rev Genet 2013 149. 2013;14(9):661–73.

217. Jonkers IH, Wijmenga C. Context-specific effects of genetic variants associated with autoimmune disease. Hum Mol Genet. 2017;26(R2):R185–92.

218. Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, Thomson BP, et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat Genet. 2010;42(6):508–14.

219. Bentham J, Morris DL, Cunninghame Graham DS, Pinder CL, Tombleson P, Behrens TW, et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. Nat Genet. 2015;47(12):1457–64.

220. Harley JB, Alarcón-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, PXK, KIAA1542 and other loci. Nat Genet. 2008;40(2):204–10.

221. Dubois PCA, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A, et al. Multiple common variants for celiac disease influencing immune gene expression. Nat Genet. 2010;42(4):295–302.

222. Miller FW, Chen W, O'Hanlon TP, Cooper RG, Vencovsky J, Rider LG, et al. Genome-wide association study identifies HLA 8.1 ancestral haplotype alleles as major genetic risk factors for myositis phenotypes. Genes Immun. 2015;16(7):470–80.

223. Hochberg MC. Updating the American College of Rheumatology. Arthritis Rheumatol. 1997;40(9):1725–34.

224. Arnett FC. The 1987 American Rheumatism Association revised criteria for the classification of rheumatoid arthritis. Arthritis Rheumatol. 1987;31(3):13–4.

225. Bohan A, Peter JB, Bowman RL. A computer-assisted analysis of 153 patients with polymyositis and dermatomyositis. Med. 1977;56:255–86.

226. Han B, Eskin E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. Am J Hum Genet. 2011;88(5):586–98.

227. Yang J, Ferreira T, Morris AP, Medland SE, Madden PAF, Heath AC, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet. 2012;44(4):369–75.

228. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88(1):76–82.

229. Chelala C, Khan A, Lemoine NR. SNPnexus: A web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. Bioinformatics. 2009;25(5):655–61.

230. Bhattacharjee S, Rajaraman P, Jacobs KB, Wheeler WA, Melin BS,

Hartge P, et al. A Subset-Based Approach Improves Power and Interpretation for the Combined Analysis of Genetic Association Studies of Heterogeneous Traits. Am J Hum Genet. 2012;90:821-35.

231. Yu W, Clyne M, Khoury MJ, Gwinn M. Phenopedia and Genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations. Bioinformatics. 2010;26(1):145–6.

232. Iotchkova V, Ritchie GRS, Geihs M, Morganella S, Min JL, Walter K, et al. GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. Nat Genet 2019 512. 2019;51(2):343–53.

233. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, et al. Defining functional DNA elements in the human genome. Proc Natl Acad Sci. 2014;111(17):6131–8.

234. Schofield EC, Carver T, Achuthan P, Freire-Pritchett P, Spivakov M, Todd JA, et al. CHiCP: a web-based tool for the integrative and interactive visualization of promoter capture Hi-C datasets. Bioinformatics. 2016;32(16):2511–3.

235. Carvalho-Silva D, Pierleoni A, Pignatelli M, Ong CK, Fumis L, Karamanis N, et al. Open Targets Platform: New developments and updates two years on. Nucleic Acids Res. 2019;47(D1):D1056–65.

236. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018;46(D1):D1074–82.

237. Mells GF, Floyd JAB, Morley KI, Cordell HJ, Franklin CS, Shin S-Y, et al. Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis. Nat Genet. 2011;43(4):329–32.

238. Cordell HJ, Han Y, Mells GF, Li Y, Hirschfield GM, Greene CS, et al. International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. Nat Commun 2015 61.

2015;6(1):1–11.

239. Senapati S, Singh S, Das M, Kumar A, Gupta R, Kumar U, et al. Genome-wide analysis of methotrexate pharmacogenomics in rheumatoid arthritis shows multiple novel risk variants and leads for TYMS regulation. Pharmacogenet Genomics. 2014;24(4):211–9.

240. Dustin ML, Cooper JA. The immunological synapse and the actin cytoskeleton: molecular hardware for T cell signaling. Nat Immunol. 2000;1(1):23–9.

241. de Vries PS, Chasman DI, Sabater-Lleal M, Chen M-H, Huffman JE, Steri M, et al. A meta-analysis of 120 246 individuals identifies 18 new loci for fibrinogen concentration. Hum Mol Genet. 2016;25(2):358–70.

242. Davalos D, Akassoglou K. Fibrinogen as a key regulator of inflammation in disease. Semin Immunopathol. 2011;34(1):43–62.

243. Araki Y, Mimura T. The Histone Modification Code in the Pathogenesis of Autoimmune Diseases. Mediators Inflamm. 2017;2017:2608605.

244. Cho JH, Feldman M. Heterogeneity of autoimmune diseases: pathophysiologic insights from genetics and implications for new therapies. Nat Med. 2015;21(7):730–8.

245. Zhernakova A, Withoff S, Wijmenga C. Clinical implications of shared genetics and pathogenesis in autoimmune diseases. Nat Rev Endocrinol 2013 911. 2013;9(11):646–59.

246. Zhernakova A, van Diemen CC, Wijmenga C. Detecting shared pathogenesis from the shared genetics of immune-related diseases. Nat Rev Genet 2009 101. 2009;10(1):43–55.

247. Kearney SJ, Delgado C, Eshleman EM, Hill KK, O'Connor BP, Lenz LL. Type I IFNs Downregulate Myeloid Cell IFN-γ Receptor by Inducing Recruitment of an Early Growth Response 3/NGFI-A Binding Protein 1 Complex That Silences ifngr1 Transcription. J Immunol. 2013;191(6):3384–

92.

248. Postal M, Vivaldo JF, Fernandez-Ruiz R, Paredes JL, Appenzeller S, Niewold TB. Type I interferon in the pathogenesis of systemic lupus erythematosus. Curr Opin Immunol. 2020;67:87–94.

249. Wright HL, Thomas HB, Moots RJ, Edwards SW. Interferon gene expression signature in rheumatoid arthritis neutrophils correlates with a good response to TNFi therapy. Rheumatology. 2015;54(1):188–93.

250. De Paepe B. Interferons as components of the complex web of reactions sustaining inflammation in idiopathic inflammatory myopathies. Cytokine. 2015;74(1):81–7.

251. Van Baal J, De Widt J, Divecha N, Van Blitterswijk WJ. Diacylglycerol kinase θ counteracts protein kinase C-mediated inactivation of the EGF receptor. Int J Biochem Cell Biol. 2012;44(11):1791–9.

252. Overbeek MJ, Boonstra A, Voskuyl AE, Vonk MC, Vonk-Noordegraaf A, van Berkel MP, et al. Platelet-derived growth factor receptor-β and epidermal growth factor receptor in pulmonary vasculature of systemic sclerosis-associated pulmonary arterial hypertension versus idiopathic pulmonary arterial hypertension and pulmonary veno-occlusive disease: a case-control study. Arthritis Res Ther 2011 132. 2011;13(2):1–13.

253. Lessard CJ, Li H, Adrianto I, Ice JA, Rasmussen A, Grundahl KM, et al. Variants at multiple loci implicated in both innate and adaptive immune responses are associated with Sjögren's syndrome. Nat Genet. 2013;45(11):1284–94.

254. T C, M Z, TRDJ R, W M. Novel insights into dendritic cells in the pathogenesis of systemic sclerosis. Clin Exp Immunol. 2020;201(1):25–33.

255. Paul P, Van Den Hoorn T, Jongsma MLM, Bakker MJ, Hengeveld R, Janssen L, et al. A Genome-wide Multidimensional RNAi Screen Reveals Pathways Controlling MHC Class II Antigen Presentation. Cell. 2011;145(2):268–83.

256. Ishaq M, Lin B-R, Bosche M, Zheng X, Yang J, Huang D, et al. LIM kinase 1 - dependent cofilin 1 pathway and actin dynamics mediate nuclear retinoid receptor function in T lymphocytes. BMC Mol Biol. 2011;12:41.

257. Duvall MG, Fuhlbrigge ME, Reilly RB, Walker KH, Kiliç A, Levy BD. Human NK Cell Cytoskeletal Dynamics and Cytotoxicity Are Regulated by LIM Kinase. J Immunol. 2020;205(3):801–10.

258. Emerging Risk Factor Collaboration, Kaptoge S, Angelantonio E, Pennells L, Wood AM, White IR, et al. C-Reactive Protein, Fibrinogen, and Cardiovascular Disease Prediction. N J Engl Med. 2012;367:1310-20.

259. Pauling JD, Christopher-Stine L. The aetiopathogenic significance, clinical relevance and therapeutic implications of vasculopathy in idiopathic inflammatory myopathy. Rheumatology. 2021;60(4):1593–607.

260. Tselios K, Urowitz MB. Cardiovascular and Pulmonary Manifestations of Systemic Lupus Erythematosus. Curr Rheumatol Rev. 2017;13(3).

261. 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. 2010;69(3):550–5.

262. Bossini-Castillo L, Broen JCA, Simeon CP, Beretta L, Vonk MC, Ortego-Centeno N, 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):638–41.

263. Sacchetti C, Bottini N. Protein Tyrosine Phosphatases in Systemic Sclerosis: Potential Pathogenic Players and Therapeutic Targets. Curr Rheumatol Reports. 2017;19(5):1–12.

264. Huang X-L, Wang Y-J, Yan J-W, Wan Y-N, Chen B, Li B-Z, et al. Role of anti-inflammatory cytokines IL-4 and IL-13 in systemic sclerosis. Inflamm Res. 2015;64(3):151–9.

265. Lewis MJ, Vyse S, Shields AM, Boeltz S, Gordon PA, Spector TD, et al.

UBE2L3 Polymorphism Amplifies NF-κB Activation and Promotes Plasma Cell Development, Linking Linear Ubiquitination to Multiple Autoimmune Diseases. Am J Hum Genet. 2015;96(2):221–34.

266. Genetech Inc. Genentech's Actemra Becomes the First Biologic Therapy Approved by the FDA for Slowing the Rate of Decline in Pulmonary Function in Adults With Systemic Sclerosis-Associated Interstitial Lung Disease, a Rare, Debilitating Condition. 2021. Available from: https://www.gene.com/media/press-releases/14897/2021-03-04/genentechs-actemra-becomes-the-first-bio

267. Andrews S. FastQC: a quality control tool for high throughput sequence data. Vol. 53, Journal of Chemical Information and Modeling. 2010.

268. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17(1):10–2.

269. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.

270. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.

271. Tang H, Peng J, Wang P, Risch NJ. Estimation of individual admixture: Analytical and study design considerations. Genet Epidemiol. 2005;28(4):289–301.

272. Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, Risch N. Estimating kinship in admixed populations. Am J Hum Genet. 2012;91(1):122–38.

273. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet. 2012;44(8):955–9.

274. Loh PR, Danecek P, Palamara PF, Fuchsberger C, Reshef YA, Finucane HK, et al. Reference-based phasing using the Haplotype Reference

Consortium panel. Nat Genet. 2016;48(11):1443–8.

275. Shabalin AA. Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012;28(10):1353–8.

276. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013;45(10):1238–43.

277. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.

278. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12(5):453–7.

279. Edgar R. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30(1):207–10.

280. Tan G, Lenhard B. TFBSTools: An R/bioconductor package for transcription factor binding site analysis. Bioinformatics. 2016;32(10):1555–6.

281. Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, Van Der Lee R, et al. JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. 2018;46(D1):D260–6.

282. Lappalainen T, Sammeth M, Friedländer MR, 'T Hoen PAC, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;501(7468):506–11.

283. Acosta-Herrera M, López-Isac E, Martín J. Towards a Better Classification and Novel Therapies Based on the Genetics of Systemic Sclerosis. Curr Rheumatol Rep. 2019;21(9):1–7.

284. Volkmann ER, Varga J. Emerging targets of disease-modifying therapy for systemic sclerosis. Nat Rev Rheumatol. 2019;15(4):208–24.

285. Hinz B, Lagares D. Evasion of apoptosis by myofibroblasts: a hallmark of fibrotic diseases. Nat Rev Rheumatol. 2020;16(1):11–31.

286. Zhernakova D V, Deelen P, Vermaat M, van Iterson M, van Galen M, Arindrarto W, et al. Identification of context-dependent expression quantitative trait loci in whole blood. Nat Genet. 2016;49(1):139–45.

287. Zweers MC, Hakim AJ, Grahame R, Schalkwijk J. Joint hypermobility syndromes: The pathophysiologic role of tenascin-X gene defects. Arthritis Rheum. 2004;50(9):2742–9.

288. Malfait F, Francomano C, Byers P, Belmont J, Berglund B, Black J, et al. The 2017 international classification of the Ehlers–Danlos syndromes. Am J Med Genet Part C Semin Med Genet. 2017;175(1):8–26.

289. Zhou X, Li H, Guo S, Wang J, Shi C, Espitia M, et al. Associations of multiple NOTCH4 exonic variants with systemic sclerosis. J Rheumatol. 2019;46(2):184–9.

290. Zhang F, Michaelson JE, Moshiach S. Tetraspanin CD151 maintains vascular stability by balancing the forces of cell adhesion and cytoskeletal tension. Blood. 2014;123(24):3843.

291. Dees C, Tomcik M, Zerr P, Akhmetshina A, Horn A, Palumbo K, et al. Notch signalling regulates fibroblast activation and collagen release in systemic sclerosis. Ann Rheum Dis. 2011;70(7):1304–10.

292. Dees C, Zerr P, Tomcik M, Beyer C, Horn A, Akhmetshina A, et al. Inhibition of Notch signaling prevents experimental fibrosis and induces regression of established fibrosis. Arthritis Rheum. 2011;63(5):1396–404.

293. James AC, Szot JO, Iyer K, Major JA, Pursglove SE, Chapman G, et al. Notch4 reveals a novel mechanism regulating Notch signal transduction. Biochim Biophys Acta - Mol Cell Res. 2014;1843(7):1272–84.

294. Pitulescu ME, Schmidt I, Giaimo BD, Antoine T, Berkenfeld F, Ferrante F, et al. Dll4 and Notch signalling couples sprouting angiogenesis and artery formation. Nat Cell Biol. 2017;19(8):915–27.

295. Limbourg FP, Takeshita K, Radtke F, Bronson RT, Chin MT, Liao JK. Essential role of endothelial Notch1 in angiogenesis. Circulation. 2005;111(14):1826–32.

296. Chen M, Daha MR, Kallenberg CGM. The complement system in systemic autoimmune disease. J Autoimmun. 2010;34(3):J276–86.

297. Kamitaki N, Sekar A, Handsaker RE, Rivera H De, Tooley K, Morris DL, et al. Complement genes contribute sex-biased vulnerability in diverse disorders. Nature. 2020;582:577-581.

298. Catrysse L, Vereecke L, Beyaert R, van Loo G. A20 in inflammation and autoimmunity. Trends Immunol. 2014;35(1):22–31.

299. Shi G, Abbott KN, Wu W, Salter RD, Keyel PA. Dnase1L3 regulates inflammasome-dependent cytokine secretion. Front Immunol. 2017;8:1–16.

300. Peña-Blanco A, García-Sáez AJ. Bax, Bak and beyond — mitochondrial performance in apoptosis. FEBS J. 2018;285(3):416–31.

301. Lagares D, Santos A, Grasberger PE, Liu F, Probst CK, Rahimi RA. Targeted Apoptosis of Myofibroblasts with the BH3 Mimetic ABT-263 Reverses Established Fibrosis. Sci Trans Med. 2017;9(420):eeal3765.

302. Liu G, Hu Y, Jin S, Zhang F, Jiang Q, Hao J. Cis-eQTLs regulate reduced LST1 gene and NCR3 gene expression and contribute to increased autoimmune disease risk. Proc Natl Acad Sci. 2016;113(42):E6321–2.

303. Fallahi P, Ruffilli I, Giuggioli D, Colaci M, Ferrari SM, Antonelli A, et al. Associations between Systemic Sclerosis and Thyroid Diseases. Front Endocrinol (Lausanne). 2017;8:266.

304. Ferrera F, Rizzi M, Sprecacenere B, Balestra P, Sessarego M, Di Carlo A, et al. AIRE gene polymorphisms in systemic sclerosis associated with autoimmune thyroiditis. Clin Immunol. 2007;122(1):13–7.

305. Gallant S, Gilkeson G. ETS transcription factors and regulation of immunity. Arch Immunol Ther Exp (Warsz). 2006;54(3):149–63.

306. Wang CY, Petryniak B, Thompson CB, Kaelin WG, Leiden JM.

Regulation of the Ets-related transcription factor Elf-1 by binding to the retinoblastoma protein. Science. 1993;260(5112):1330–5.

307. Shores EW, Love PE. TCR ζ chain in T cell development and selection. Curr Opin Immunol. 1997;9(3):380–9.

308. Grant PA, Thompson CB, Pettersson S. IgM receptor-mediated transactivation of the IgH 3′ enhancer couples a novel Elf-1-AP-1 protein complex to the developmental control of enhancer function. EMBO J. 1995;14(18):4501–13.

309. Thalayasingam N, Nair N, Skelton AJ, Massey J, Anderson AE, Clark AD, et al. CD4+ and B Lymphocyte Expression Quantitative Traits at Rheumatoid Arthritis Risk Loci in Patients With Untreated Early Arthritis: Implications for Causal Gene Identification. Arthritis Rheumatol. 2018;70(3):361–70.

310. James T, Lindén M, Morikawa H, Fernandes SJ, Ruhrmann S, Huss M, et al. Impact of genetic risk loci for multiple sclerosis on expression of proximal genes in patients. Hum Mol Genet. 2018;27(5):912–28.

311. Odhams CA, Cortini A, Chen L, Roberts AL, Viñuela A, Buil A, et al. Mapping eQTLs with RNA-seq reveals novel susceptibility genes, non-coding RNAs and alternative-splicing events in systemic lupus erythematosus. Hum Mol Genet. 2017;26(5):1003–17.

312. Newman JRB, Conesa A, Mika M, New FN, Onengut-Gumuscu S, Atkinson MA, et al. Disease-specific biases in alternative splicing and tissue-specific dysregulation revealed by multitissue profiling of lymphocyte gene expression in type 1 diabetes. Genome Res. 2017;27(11):1807–15.

313. Skaug B, Khanna D, Swindell WR, Hinchcliff ME, Frech TM, Steen VD, et al. Global skin gene expression analysis of early diffuse cutaneous systemic sclerosis shows a prominent innate and adaptive inflammatory profile. Ann Rheum Dis. 2019;1–8.

314. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ

preprocessor. Bioinformatics. 2018;34(17):i884–90.

315. Wingett S, Ewels P, Furlan-Magaril M, Nagano T, Schoenfelder S, Fraser P, et al. HiCUP: pipeline for mapping and processing Hi-C data. F1000Research. 2015;4.

316. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012 94. 2012;9(4):357–9.

317. Cairns J, Freire-Pritchett P, Wingett SW, Várnai C, Dimond A, Plagnol V, et al. CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. Genome Biol 2016 171. 2016;17(1):1–17.

318. Cairns J, Orchard WR, Malysheva V, Spivakov M. Chicdiff: a computational pipeline for detecting differential chromosomal interactions in Capture Hi-C data. Bioinformatics. 2019;35(22):4764–6.

319. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166–9.

320. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists. Nucleic Acids Res. 2019;47(W1):W191–8.

321. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012;9(3):215–6.

322. Boix CA, James BT, Park YP, Meuleman W, Kellis M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. Nature. 2021;590(7845):300–7.

323. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for Computing and Annotating Genomic Ranges. PLOS Comput Biol. 2013;9(8):e1003118.

324. Zhou X, Lowdon RF, Li D, Lawson HA, Madden PAF, Costello JF, et al. Exploring long-range genome interactions using the WashU Epigenome Browser. Nat Methods. 2013;10(5):375–6.

325. Bonifazi M, Tramacere I, Pomponio G, Gabrielli B, Avvedimento E

V., La Vecchia C, et al. Systemic sclerosis (scleroderma) and cancer risk: systematic review and meta-analysis of observational studies. Rheumatology. 2013;52(1):143–54.

326. Mendoza FA, Piera-Velazquez S, Jimenez SA. Tyrosine kinases in the pathogenesis of tissue fibrosis in systemic sclerosis and potential therapeutic role of their inhibition. Transl Res. 2021;231:139–58.

327. Jeng MY, Mumbach MR, Granja JM, Satpathy AT, Chang HY, Chang ALS. Enhancer Connectome Nominates Target Genes of Inherited Risk Variants from Inflammatory Skin Disorders. J Invest Dermatol. 2019 Mar 1;139(3):605–14.

328. Moser B. CXCR5, the Defining Marker for Follicular B Helper T (TFH) Cells. Front Immunol. 2015;0:296.

329. Ricard L, Jachiet V, Malard F, Ye Y, Stocker N, Rivière S, et al. Circulating follicular helper T cells are increased in systemic sclerosis and promote plasmablast differentiation through the IL-21 pathway which can be inhibited by ruxolitinib. Ann Rheum Dis. 2019;78(4):539–50.

330. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012;491(7422):119–24.

331. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell. 2016;167(5):1415-1429.e19.

332. Wang Q, Wang C, Li N, Liu X, Ren W, Wang Q, et al. Condensin Smc4 promotes inflammatory innate immune response by epigenetically enhancing NEMO transcription. J Autoimmun. 2018;92:67–76.

333. Hua P, Badat M, Hanssen LLP, Hentges LD, Crump N, Downes DJ, et al. Defining genome architecture at base-pair resolution. Nature. 2021 Jun 9;595(7865):125–9.

334. Salem S, Salem D, Gros P. Role of IRF8 in immune cells functions,

protection against infections, and susceptibility to inflammatory diseases. Hum Genet. 2020;139(6):707–21.

335. Crosslin DR, McDavid A, Weston N, Zheng X, Hart E, de Andrade M, et al. Genetic variation associated with circulating monocyte count in the eMERGE Network. Hum Mol Genet. 2013;22(10):2119–27.

336. Ototake Y, Yamaguchi Y, Asami M, Komitsu N, Akita A, Watanabe T, et al. Downregulated IRF8 in Monocytes and Macrophages of Patients with Systemic Sclerosis May Aggravate the Fibrotic Phenotype. J Invest Dermatol. 2021;141(8):1954–63.

337. Murakami K, Sasaki H, Nishiyama A, Kurotaki D, Kawase W, Ban T, et al. A RUNX–CBFβ-driven enhancer directs the Irf8 dose-dependent lineage choice between DCs and monocytes. Nat Immunol. 2021;22(3):301–11.

338. Irving BA, Chan AC, Weiss A. Functional characterization of a signal transducing motif present in the T cell antigen receptor zeta chain. J Exp Med. 1993;177(4):1093–103.

339. O'Shea JJ, Lahesmaa R, Vahedi G, Laurence A, Kanno Y. Genomic views of STAT function in CD4+ T helper cell differentiation. Nat Rev Immunol. 2011;11(4):239–50.

340. Tarbell E, Jiang K, Hennon TR, Holmes L, Williams S, Fu Y, et al. CD4+ T cells from children with active juvenile idiopathic arthritis show altered chromatin features associated with transcriptional abnormalities. Sci Reports. 2021;11(1):1–14.

341. Greenwald WW, Li H, Benaglio P, Jakubosky D, Matsui H, Schmitt A, et al. Subtle changes in chromatin loop contact propensity are associated with differential gene regulation and expression. Nat Commun. 2019;10(1):1–17.

342. Mohammadi S, Davila-Velderrain J, Kellis M. Reconstruction of Cell-type-Specific Interactomes at Single-Cell Resolution. Cell Syst. 2019;9(6):559-568.e4.

343. Ingwersen J, Aktas O, Kuery P, Kieseier B, Boyko A, Hartung HP.

Fingolimod in multiple sclerosis: Mechanisms of action and clinical efficacy. Clin Immunol. 2012;142(1):15–24.

344. Chatterjee S, Ahituv N. Gene Regulatory Elements, Major Drivers of Human Disease. Annu Rev Genomics Hum Genet. 2017;18:45–63.

345. Xie Y, Su N, Yang J, Tan Q, Huang S, Jin M, et al. FGF/FGFR signaling in health and disease. Signal Transduct Target Ther. 2020;5(1):1–38.

346. Pepelyayeva Y, Amalfitano A. The role of ERAP1 in autoinflammation and autoimmunity. Hum Immunol. 2019;80(5):302–9.

347. Reeves E, Islam Y, James E. ERAP1: a potential therapeutic target for a myriad of diseases. Expert Opin Ther Targets. 2020;24(6):535–44.

348. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 2017;169(7):1177–86.

349. Liu X, Li YI, Pritchard JK. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. Cell. 2019;177(4):1022-1034.e6.

350. Shifrut E, Carnevale J, Tobin V, Roth TL, Woo JM, Bui CT, et al. Genome-wide CRISPR Screens in Primary Human T Cells Reveal Key Regulators of Immune Function. Cell. 2018;175(7):1958-1971.e15.

351. Thynn HN, Chen XF, Hu WX, Duan YY, Zhu DL, Chen H, et al. An Allele-Specific Functional SNP Associated with Two Systemic Autoimmune Diseases Modulates IRF5 Expression by Long-Range Chromatin Loop Formation. J Invest Dermatol. 2020;140(2):348-360.e11.