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


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

Abstract

The aim of this study was to determine, through a genome-wide association study (GWAS), the genetic components contributing to different clinical sub-phenotypes of systemic sclerosis (SSc). We considered limited (lcSSc) and diffuse (dcSSc) cutaneous involvement, and the relationships with presence of the SSc-specific auto-antibodies, anti-centromere (ACA), and anti-topoisomerase I (ATA). Four GWAS cohorts, comprising 2,296 SSc patients and 5,171 healthy controls, were included. The first employed, comprising 1,361 SSc patients and 3,586 controls, was used to identify genetic associations with different cutaneous involvement. Genome-wide significant SNPs (P ≤ 5 × 10−8) were found at the HLA-DPA1/B1 locus with ACA (OR = 0.75). Also, rs12540874 in GRB10 gene (P = 1.27 × 10−6; OR = 1.15) and rs11047102 in SOX5 gene (P = 1.39 × 10−7; OR = 1.36) showed a suggestive association with lcSSc and ACA subgroups respectively. In the HLA region, we observed highly associated allelic combinations in the HLA-DQB1 locus with ACA (P = 1.79 × 10−6; OR = 2.48), in the HLA-DPA1/B1 loci with ATA (P = 4.57 × 10−7; OR = 8.84), and in NOTCH4 with ACA P = 8.84 × 10−21; OR = 0.55) and ATA (P = 1.14 × 10−8, OR = 6.84).
Introduction

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

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

It has been observed that certain SSc clinical features and the presence of disease specific auto-antibodies vary in different countries and ethnicities [20]. This fact supports the likelihood that genetic factors may influence the different clinical features of the disease and auto-antibody subsets [19]. Furthermore, the affected members within multigene SSc families tend to be concordant for SSc-specific auto-antibodies and HLA haplotypes, thus, providing further evidence for a genetic basis for auto-antibody expression in SSc [21]. Moreover, several studies have reported that certain SSc genetic risk factors correlate with specific clinical subsets of the disease or SSc-related auto-antibodies [4,12,22,23].

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

Results

First, the genetic associations were tested in each of the four subgroups considered for this study (lcSSc, dcSSc, ACA positive and ATA positive) by the means of χ² tests in the GWAS data (individuals from the United States, Spain, Germany and The Netherlands), correcting the P values for the genomric inflation factor λ of each subgroup (Figures S1, S2, S3, S4 and Tables S1, S2, S3, S4). We found a total of eighteen novel non-HLA loci associated in these subgroups with a P value lower than 1×10⁻³⁷, seven in the lcSSc subtype, five in the dcSSc subtype, two in ACA positives and four in ATA positives. Next, we proceeded to replicate these associations in nine independent cohorts (from US, Spain, Germany, The Netherlands, Belgium, Italy, Sweden, United Kingdom and Norway). The statistically significant results observed in the replication step are shown in Table 1. The complete set of data is shown in Tables S1, S2, S3 S4.

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

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

**Clinical Manifestations**

In the lcSSc subtype, seven non-HLA novel loci were identified as susceptibility markers in the GWAS data (Table S1 and Figure S1). Two out of the seven genetic markers showed evidence of association in the replication cohorts: rs11642973 near the IRF8 gene (lcSSc $P = 2.32 \times 10^{-12}$, OR = 0.75 [0.69-0.81]) at the GWAS level of significance and rs12540874 in the GRB10 gene (lcSSc $P = 1.27 \times 10^{-6}$, OR = 1.15 [1.09-1.22]) at the suggestive level of significance (Figure 1, Table 1 and Table S1).

Regarding the dcSSc subtype, five non-HLA loci were found to be associated in the GWAS cohorts (Table S2 and Figure S2). Upon analyzing these five SNPs in the replication cohorts, we could only replicate the association of rs11171747 in the RPLA1/EST7 locus (overall dcSSc $P = 5.99 \times 10^{-6}$, OR = 1.23 [1.14-1.33]) (Figure 1, Table 1 and Table S2). However, the association found in this locus was heterogeneous among cohorts (Breslow-Day $P = 5.32 \times 10^{-5}$).

**Auto-Antibodies**

The observed associations in the ACA positive subgroup and lcSSc were difficult to differentiate because of substantial overlap between these two disease subgroups. In the GWAS cohorts, SNPs in IL12RB2 and RUNX1 genes were identified as novel non-HLA loci associated with SSC patients positive for ACA antibodies (Table S3 and Figure S3). However, none of these associations could be confirmed at the replication stage. Interestingly, the SNP rs11047102 of the RUNX1 gene, which was selected for replication (ACA $P = 2.32 \times 10^{-12}$, OR = 0.75 [0.69-0.81]) at the GWAS level of significance and rs9275390 (ACA+ $P = 2.62 \times 10^{-5}$, OR = 2.30 [2.13-2.67]). The TC allele combination (both risk alleles) showed a high association in the ACA positive subgroup (ACA+ $P = 9.81 \times 10^{-12}$, OR = 2.40 [2.22-2.77]), being present in 45.3% of the ACA positive patients compared to 25.1% of the controls (Table S3).

Regarding the ATA positive subgroup, we also observed evidence of independent association in the HLA region (Table 2 and Figure 1, Table S7). We found three associations in the HLA class II region: rs3129763 in the HLA-DQA1/DRB1 loci (ATA+ $P = 1.47 \times 10^{-11}$, OR = 1.65 [1.42-1.91]) and four associated SNPs in the HLA-DPA1/DPB1 region (highest association at rs987870, ATA+ $P = 2.42 \times 10^{-20}$, OR = 2.09 [1.78-2.45]). The combination of three risk alleles in the DPA1/DPB1 locus, (CAC ATA+ $P = 1.27 \times 10^{-7}$, OR = 8.84 [6.72-11.63]) of the SNPs rs987870, rs1315021 and rs6901221 respectively was present in 10.6% of the ATA positive SSC patients compared to only 1.3% of the controls (Table S3). In addition, in the HLA class III region, the NOTCH4 gene was associated with the presence of ACA (rs443198, ACA+ $P = 8.84 \times 10^{-21}$, OR = 0.55 [0.49-0.63] and ATA (rs9296015, ATA+ $P = 1.14 \times 10^{-8}$, OR = 0.54 [0.44-0.67]), independently of the HLA class II associations (Table 2 and Tables S6, S7). Interestingly, SNP rs9296015 had an opposite effect size in ACA and ATA subgroup, being exclusively associated in the ATA subgroup. These two SNPs were not in LD in Caucasian populations either from the HapMap project ($r^2 = 0.05$ in CEU and $r^2 = 0.03$ in TSI) or our cohorts ($r^2 = 0.11$ in Spanish, $r^2 = 0.00$ in German, $r^2 = 0.01$ in Dutch and $r^2 = 0.01$ in US), pointing to independent associations in the NOTCH4 gene with both ACA and ATA positive subgroups. All the associations ORs found in the HLA region were consistent among the four GWAS cohorts (Tables S8, S9).

**Previously Described Genetic Associations**

We wanted to investigate previously reported associations with subphenotypes or overall disease, such as CD247, TNFSF4, STAT4, BANK1, IRF5 and BLK in the present study’s GWAS cohorts, to further establish them as SSC (or its subphenotypes) susceptibility loci. Table S10 shows the analysis of the SNPs in the previously mentioned genes which were present in our GWAS combined panel. As expected, association previously found in these six genes was replicated. Interestingly associations previously described to be confined to one of the SSC subgroups were also replicated as in the cases of TNFSF4 and lcsSc (lcsSc $P = 7.70 \times 10^{-4}$, OR = 1.18 [1.03-1.31]), STAT4 and lcsSc (lcsSc $P = 7.70 \times 10^{-6}$, OR = 1.31 [1.19-1.48]), BANK1 and dcSSc (dcSSc $P = 0.0103$, OR = 0.85 [0.75-0.96]) and BLK and ACA+ (ACA+ $P = 1.43 \times 10^{-5}$, OR = 1.27 [1.12-1.44]). Furthermore association of CD247 with SSC was more strongly represented in the lcsSc subgroup than the others (lcsSc $P = 2.66 \times 10^{-6}$, OR = 0.81 [0.75-0.89]), although evidence of association was also
found in the other subgroups. Similarly, the association found in IRF5 was stronger in lcSSc (lcSSc P = 1.64 × 10⁻², OR = 1.50 [1.32–1.69]), although association was also found in the dcSSc, ACA+ and ATA+ subgroups.

**Discussion**

Systemic sclerosis (SSc) is a rare, severe, complex and heterogeneous rheumatic disease. Multiple lines of evidence suggest that genetic factors may underlie not only SSc susceptibility but also the predisposition to develop specific clinical phenotypes such as lcSSc, dcSSc subtypes and the presence of SSc-specific auto-antibodies. The discovery of genetic variants associated with specific clinical manifestations of the disease will lead to new insights regarding pathogenesis and may open novel avenues of therapy that can be targeted to specific subsets.

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

---

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

<table>
<thead>
<tr>
<th>SSc Subphenotype</th>
<th>Chr.</th>
<th>Gene</th>
<th>SNP</th>
<th>Base Pair</th>
<th>Location</th>
<th>Change</th>
<th>Stage</th>
<th>N (case/control)</th>
<th>MAF (case/control)</th>
<th>P value</th>
<th>OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>lcSSc</td>
<td>7p12.1</td>
<td>GRB10</td>
<td>rs12540874</td>
<td>50,632,416</td>
<td>Intronic</td>
<td>G/A</td>
<td>GWAS</td>
<td>1400/5172</td>
<td>0.461/0.409</td>
<td>3.00 × 10⁻⁶</td>
<td>1.23 (1.13–1.34)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>1960/4971</td>
<td>0.416/0.395</td>
<td>3.07 × 10⁻²</td>
<td>1.09 (1.01–1.18)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combined</td>
<td>3360/10143</td>
<td>0.435/0.403</td>
<td>1.27 × 10⁻⁵</td>
<td>1.15 (1.09–1.22)</td>
</tr>
<tr>
<td>16q24.1</td>
<td></td>
<td>IRF8</td>
<td>rs11642873</td>
<td>84,549,206</td>
<td>Intergenic</td>
<td>C/A</td>
<td>GWAS</td>
<td>1400/5172</td>
<td>0.146/0.197</td>
<td>1.39 × 10⁻⁷</td>
<td>0.72 (0.64–0.81)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>1960/4971</td>
<td>0.143/0.186</td>
<td>6.88 × 10⁻⁶</td>
<td>0.77 (0.70–0.87)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combined</td>
<td>3360/10143</td>
<td>0.144/0.192</td>
<td>2.32 × 10⁻¹²</td>
<td>0.75 (0.69–0.81)</td>
</tr>
<tr>
<td>dcSSc</td>
<td>12q13.2</td>
<td>RPL41/ESYT1*</td>
<td>rs11171747</td>
<td>54,804,675</td>
<td>Upstream</td>
<td>G/T</td>
<td>GWAS</td>
<td>740/5172</td>
<td>0.446/0.384</td>
<td>2.19 × 10⁻⁶</td>
<td>1.31 (1.01–1.29)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>959/4971</td>
<td>0.408/0.372</td>
<td>3.49 × 10⁻³</td>
<td>1.16 (1.15–1.71)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combined</td>
<td>1699/10143</td>
<td>0.425/0.379</td>
<td>5.99 × 10⁻⁸</td>
<td>1.23 (1.10–1.50)</td>
</tr>
<tr>
<td>ACA+</td>
<td>12p12.1</td>
<td>SOX5</td>
<td>rs11047102</td>
<td>23,837,413</td>
<td>Intronic</td>
<td>T/C</td>
<td>GWAS</td>
<td>761/5172</td>
<td>0.132/0.096</td>
<td>1.03 × 10⁻⁵</td>
<td>1.47 (1.24–1.73)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>1030/4971</td>
<td>0.123/0.102</td>
<td>2.91 × 10⁻³</td>
<td>1.27 (1.09–1.48)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Combined</td>
<td>1791/10143</td>
<td>0.127/0.099</td>
<td>1.39 × 10⁻⁷</td>
<td>1.36 (1.21–1.52)</td>
</tr>
</tbody>
</table>

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

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

doi:10.1371/journal.pgen.1002178.t001

---

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

doi:10.1371/journal.pgen.1002178.g001
subphenotypes through an analysis of our previous genome-wide association study (GWAS) data stratified for these disease subphenotypes, together with a large, new replication study.

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

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

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

The most prominent SSc specific auto-antibodies, ACA and ATA, are associated with the lcSSc and dcSSc clinical subsets, respectively [19]. The lcSSc subtype greatly overlaps with the ACA positive subgroup of patients (almost all ACA positive patients belonged to the lcSSc subtype). Similarly, the dcSSc subtype overlaps with the ATA positive group of patients. Therefore, it is difficult to determine whether some of the observed associations specifically belonged to one of the four subgroups. Such is the case of the association found with the SOX5 gene. In the GWAS data, SOX5 was associated with lcSSc as well as with the ACA positive subgroup, although the association with the lcSSc subtype was stronger than that in the ACA positive subgroup. Upon completion of the replication study with the resultant increase in statistical power, we were able to determine that the SOX5 gene was indeed a risk factor for the ACA positive group at the genome wide significance level, but not for lcSSc. The SOX5 gene encodes a member of the SOX (SRY-related HMG-box) family of transcription factors involved in the regulation of embryonic development, in the determination of cell fate, as well as in chondrogenesis [38].

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

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

In dcSSc patients, the only observed genome wide significant association was with the RPL41/ESIT1 locus, although this association was heterogeneous among the investigated populations, probably due to lower statistical power in this smaller group. Three genes are relevant to this locus: RPL41, a ribosomal protein not considered to be related to the immune system; ZC3H10, a zinc finger protein related to tumour growth; and ESIT1, a synaptotagmin-like protein of unknown function. Although none of these genes has a suggestive role in the pathogenesis of SSc a priori, further studies are needed to investigate this intriguing finding.

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

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

Figure 2. Manhattan plot showing the -log10 of the Mantel-Haenszel $P$ value of all 1,112 SNPs in HLA region for the GWAS cohorts comprising 2,296 cases and 5,171 controls. Associations for the whole SSc set are in black, while associations in ACA (760 cases) and ATA (447 cases) positive subgroups are in orange and red, respectively. Loci which were independently associated according to conditional logistic regression analysis are highlighted in grey.

doi:10.1371/journal.pgen.1002178.g002
Table 2. Independent associations identified in the HLA region with the ACA and ATA positive subgroups.

<table>
<thead>
<tr>
<th>SNP Selection for Replication</th>
<th>Gene</th>
<th>Location</th>
<th>Change control</th>
<th>MAF (ACA)</th>
<th>OR (CI 95%)</th>
<th>P*</th>
<th>MAF (ATA)</th>
<th>OR (CI 95%)</th>
<th>P*</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs443198</td>
<td>NOTCH4</td>
<td>Intergenic C/T</td>
<td>0.314/0.442/0.492</td>
<td>1.99</td>
<td>0.045</td>
<td>0.002</td>
<td>0.62</td>
<td>0.52</td>
<td>0.002</td>
</tr>
<tr>
<td>rs6457617</td>
<td>HLA-DQB1</td>
<td>Intergenic C/T</td>
<td>0.544/0.177/0.253</td>
<td>2.61</td>
<td>0.012</td>
<td>0.001</td>
<td>0.54</td>
<td>0.44</td>
<td>0.012</td>
</tr>
<tr>
<td>rs9275390</td>
<td>HLA-DRA</td>
<td>Intron G/A</td>
<td>0.430/0.631/0.440</td>
<td>0.2725</td>
<td>0.867</td>
<td>0.867</td>
<td>0.54</td>
<td>0.44</td>
<td>0.012</td>
</tr>
<tr>
<td>rs3129882</td>
<td>HLA-DPA1/DPB1</td>
<td>Intron C/T</td>
<td>0.139/0.270/0.146</td>
<td>0.1725</td>
<td>0.525</td>
<td>0.525</td>
<td>0.54</td>
<td>0.44</td>
<td>0.012</td>
</tr>
<tr>
<td>rs3135021</td>
<td>HLA-DPA1/DPB1</td>
<td>Intron C/A</td>
<td>0.190/0.223/0.157</td>
<td>2.98</td>
<td>0.002</td>
<td>0.002</td>
<td>0.54</td>
<td>0.44</td>
<td>0.012</td>
</tr>
<tr>
<td>rs6901221</td>
<td>HLA-DPA1/DPB1</td>
<td>Intron C/T</td>
<td>0.190/0.223/0.157</td>
<td>2.98</td>
<td>0.002</td>
<td>0.002</td>
<td>0.54</td>
<td>0.44</td>
<td>0.012</td>
</tr>
<tr>
<td>rs9287530</td>
<td>HLA-DPA1/DPB1</td>
<td>Intron C/T</td>
<td>0.190/0.223/0.157</td>
<td>2.98</td>
<td>0.002</td>
<td>0.002</td>
<td>0.54</td>
<td>0.44</td>
<td>0.012</td>
</tr>
<tr>
<td>rs3100767</td>
<td>HLA-DPA1/DPB1</td>
<td>Intron C/A</td>
<td>0.190/0.223/0.157</td>
<td>2.98</td>
<td>0.002</td>
<td>0.002</td>
<td>0.54</td>
<td>0.44</td>
<td>0.012</td>
</tr>
</tbody>
</table>

Sample size for the ACA subgroup was 761, and for ATA was 497, while the sample size for the controls was 5172.

Unadjusted P values are logistic regression analysis adjusted for all other SNPs in the same region and the same subphenotype.
overlapping exists between lcSSc and ACA+ subgroups, and dcSSc and ATA+ subgroups; we wanted to assess whether association found in overlapped groups belonged to a subtype or an auto-antibody positive group. With that purpose we selected SNPs from the GWAS data based on the following criteria:

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

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

### Genotyping

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

SNPs selected for the replication phase were genotyped in the replication cohorts using Applied Biosystems’ TaqMan SNP assays on ABI Prism 7900 HT real-time thermocyclers. Markers with call rates of 95% or less were excluded, as were markers whose allele distributions deviated strongly from Hardy-Weinberg (HW) equilibrium in controls ($P < 10^{-5}$).

### Data Imputation

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

### Statistical Analysis

Data in the SSc GWAS cohorts was filtered as follows: Using Plink, we identified and excluded pairs of genetically related subjects or duplicates and excluded the genetic-pair members with lower call rates. To identify individuals who might have non-western European ancestry, we merged our case and control data with the data from the HapMap Project (60 western European (CEU), 60 Nigerian (YRI), 90 Japanese (JPT) and 90 Han Chinese (CHB) samples). We used principal component analysis as implemented in HelixTree (see Text S2), plotting the first two principal components for each individual. All individuals who did not cluster with the main CEU cluster (defined as deviating more than 4 standard deviations from the cluster centroids) were excluded from subsequent analyses. Additionally, we excluded individuals with low call rates (11 individuals from the US group, 24 from the Spanish, 1 from the German and 1 from the Dutch), relatedness (50 from the US group, 2 from the Spanish, 1 from the German and 1 from the Dutch), non-European ancestry (42 from the US group, 5 from the Spanish, 6 from the German and 4 from the Dutch) and inconsistent gender (83 from the US group, 2 from the Spanish, 2 from the German and 2 from the Dutch). Then we filtered for SNP quality, removing SNPs with a genotyping success call rate $< 98\%$ and those showing MAF $< 1\%$. Deviation of the genotype frequencies in the controls from those expected under Hardy-Weinberg equilibrium was assessed by a $\chi^2$ test or Fisher’s exact test when an expected cell count was $< 5$. SNPs strongly deviating from Hardy-Weinberg equilibrium ($P < 10^{-5}$) were
eliminated from the study. For the combined analysis of the four datasets, the same quality controls per individual and per SNP were applied with the exception of the Hardy-Weinberg equilibrium (HWE) requirement. The genotyping success call rate on the merged dataset after all these quality filters were applied was 99.03% in the GWAS cohorts.

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

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

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

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

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

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

Web Resources
Plink software:
http://pngu.mgh.harvard.edu/purcell/plink/
SVS HelixTree software:
http://www.goldenhelix.com/SNP_Variation/HelixTree/index.html
Stata software:
http://www.stata.com/
R Statistical Package:
http://www.r-project.org/
Haploview:
http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/haplovew

Supporting Information

Figure S1  Manhattan plot and QQ plot showing the -log10 of the Mantel-Haenszel $P$ value of all 279,621 SNPs in the lcSSc individuals of the GWAS cohorts comprising 1,400 cases and 5,171 controls. All $P$ values are GC corrected, and $\lambda$ was 1.058. (TIF)

Figure S2  Manhattan plot and QQ plot showing the -log10 of the Mantel-Haenszel $P$ value of all 279,621 SNPs in the dcSSc individuals of the GWAS cohorts comprising 740 cases and 5,171 controls. All $P$ values are GC corrected, and $\lambda$ was 1.0394. (TIF)

Figure S3  Manhattan plot and QQ plot showing the -log10 of the Mantel-Haenszel $P$ value of all 279,621 SNPs in the ACA positive individuals of the GWAS cohorts comprising 761 cases and 5,171 controls. All $P$ values are GC corrected, and $\lambda$ was 1.050. (TIF)

Figure S4  Manhattan plot and QQ plot showing the -log10 of the Mantel-Haenszel $P$ value of all 279,621 SNPs in the ATA positive individuals of the GWAS cohorts comprising 447 cases and 5,171 controls. All $P$ values are GC corrected, and $\lambda$ was 1.061. (TIF)

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

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

Table S2  Analysis for GWAS cohorts, replication cohorts and combined analysis for all non-HLA, non-previously described associations with dcSSc subtype of the disease. $P$ values for GWAS cohorts are Mantel-Haenszel meta-analysis GC corrected according to the set $\lambda$, and in the replication and combined analysis Mantel-Haenszel meta-analysis $P$ value. $\lambda$ $P$ value for the totality of the SSc patients, in the case of GWAS cohorts GC corrected according to the set $\lambda$, and in replication and combined analysis Mantel-Haenszel meta-analysis $P$ value. Association in rs11177147 had a significant BD $P$ value, thus making them heterogenic associations among populations. (DOC)

Table S3  Analysis for GWAS cohorts, replication cohorts and combined analysis for all non-HLA, non-previously described...
associations with ACA positive subgroup of the disease. †P values for GWAS cohorts are Mantel-Haenszel meta-analysis GC corrected according to the set λ and in the replication and combined analysis Mantel-Haenszel meta-analysis P value. ‡P value for the totality of the SSc patients, in the case of GWAS cohorts GC corrected according to the set λ, and in replication and combined analysis Mantel-Haenszel meta-analysis P value. *Association in rs3790567 had a significant BD P value, thus making them heterogeneous associations among populations.

(DOC)

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

(DOC)

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

(DOC)

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

(DOC)

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

(DOC)

Table S8 Independent associations found in the HLA region in the ACA positive subgroup of patients in the separate four GWAS cohorts. †Uncorrected χ² P value of each separated cohort.

(DOC)

Table S9 Independent associations found in the HLA region in the ATA positive subgroup of patients in the separate four GWAS cohorts. †Uncorrected χ² P value of each separated cohort.

(DOC)

Table S10 Previously described genetic associations with SSc subphenotypes which were present in the present study’s GWAS panel of SNPs. A total of 2,296 SSc cases and 5,172 controls were included in this analysis. The SSc cases included 1,400 lcSSc individuals, 740 dcSSc individuals, 761 ACA+ individuals and 447 ATA+ individuals. Best P value in each subgroup for each SNP is in bold. Chr. Chromosome. † Uncorrected Mantel-Haenszel Meta-analysis P value of the four GWAS cohorts.

(DOC)

Table S11 Composition and size of all the populations used in the study for the considered features of the disease.

(DOC)

Text S1 Members of the Spanish Scleroderma Group.

(DOC)

Text S2 URLs. Internet Uniform Resource Locator (URL) for each of the software packages used in this study.

(DOC)

Author Contributions


(DOC)

References