

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

Timothy R D J Radstake<sup>1,38</sup>, Olga Gorlova<sup>2,38</sup>, Blanca Rueda<sup>3,38</sup>, Jose-Ezequiel Martin<sup>3,38</sup>, Behrooz Z Alizadeh<sup>4</sup>, Rogelio Palomino-Morales<sup>3</sup>, Marieke J Coenen<sup>5</sup>, Madelon C Vonk<sup>1</sup>, Alexandre E Voskuyl<sup>6</sup>, Annemie J Scheurwegh<sup>7</sup>, Jasper C Broen<sup>1</sup>, Piet L C M van Riel<sup>1</sup>, Ruben van 't Slot<sup>4</sup>, Annet Italiaander<sup>4</sup>, Roel A Ophoff<sup>4,8</sup>, Gabriela Riemekasten<sup>9</sup>, Nico Hunzelmann<sup>10</sup>, Carmen P Simeon<sup>11</sup>, Norberto Ortego-Centeno<sup>12</sup>, Miguel A González-Gay<sup>13</sup>, María F González-Escribano<sup>14</sup>, Spanish Scleroderma Group<sup>37</sup>, Paolo Airo<sup>15</sup>, Jaap van Laar<sup>16</sup>, Ariane Herrick<sup>17</sup>, Jane Worthington<sup>17</sup>, Roger Hesselstrand<sup>18</sup>, Vanessa Smith<sup>19</sup>, Filip de Keyser<sup>19</sup>, Fredric Houssiau<sup>20</sup>, Meng May Chee<sup>21</sup>, Rajan Madhok<sup>21</sup>, Paul Shiels<sup>21</sup>, Rene Westhovens<sup>22</sup>, Alexander Kreuter<sup>23</sup>, Hans Kiener<sup>24</sup>, Elfride de Baere<sup>25</sup>, Torsten Witte<sup>26</sup>, Leonid Padykov<sup>27</sup>, Lars Klareskog<sup>27</sup>, Lorenzo Beretta<sup>28</sup>, Rafaella Scorza<sup>28</sup>, Benedicte A Lie<sup>29</sup>, Anna-Maria Hoffmann-Vold<sup>30</sup>, Patricia Carreira<sup>13,31</sup>, John Varga<sup>32</sup>, Monique Hinchcliff<sup>32</sup>, Peter K Gregersen<sup>33</sup>, Annette T Lee<sup>33</sup>, Jun Ying<sup>2</sup>, Younghun Han<sup>2</sup>, Shih-Feng Weng<sup>2</sup>, Christopher I Amos<sup>2</sup>, Fredrick M Wigley<sup>34</sup>, Laura Hummers<sup>34</sup>, J Lee Nelson<sup>35</sup>, Sandeep K Agarwal<sup>36</sup>, Shervin Assasi<sup>36</sup>, Pravitt Gourh<sup>36</sup>, Filemon K Tan<sup>36</sup>, Bobby P C Koeleman<sup>4,38</sup>, Frank C Arnett<sup>36,38</sup>, Javier Martin<sup>3,38</sup> & Maureen D Mayes<sup>36,38</sup>

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

SSc is a profoundly disabling autoimmune disease characterized by vascular damage, altered immune responses and abnormal fibrosis of skin and internal organs leading to premature death in affected individuals<sup>1</sup>. The etiology of SSc is complex and poorly understood, but as with most autoimmune conditions, it is widely accepted that environmental genetic factors contribute to disease risk. Data from familial, twin and ethnicity studies support the relevance of the genetic component in SSc etiology<sup>2</sup>. Previous studies aimed at dissecting the genetic factors underlying SSc genetic susceptibility have used the candidate gene association study approach<sup>3</sup>. In spite of several years of research, this strategy has yielded a very limited characterization of SSc genetic risk factors. Except for the major histocompatibility complex (MHC) genes, which are relevant genetic markers for SSc across populations, few other loci outside the human leukocyte antigen (HLA) region have demonstrated strong and reproducible associations with SSc

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

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

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

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

susceptibility<sup>3,4</sup>. Only very recently have large case-control association studies identified *STAT4* and *IRF5* as genetic factors contributing to SSc susceptibility<sup>5–8</sup>. As for other complex genetic disorders, it is expected that several genetic markers contribute to SSc predisposition with modest effects and therefore large sample sizes are required to detect new disease-associated loci<sup>9</sup>.

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

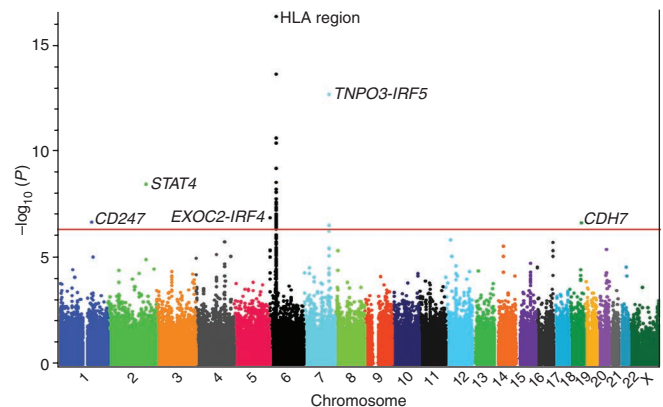
After rigorous genotyping quality-control filters, a total of 279,621 SNPs shared between the four case-control series were extracted for analysis (Supplementary Table 1).

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

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

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

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



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

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

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

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

should be further investigated. In contrast, because of the high minor allele frequency (MAF) of the rs4959270 polymorphism in the *EXOC2-IRF4* region, great heterogeneity of the association was observed in the replication cohorts (Supplementary Fig. 1). These findings are concordant with previous GWAS studies in which great population allelic heterogeneity has been reported for *EXOC2-IRF4* genetic variants leading to false positive disease associations, as may have occurred in our screening phase<sup>14</sup>. Notably, the newly identified SSc susceptibility locus, *CD247*, encodes a protein that participates in the regulation of immune response and thus could have a role in SSc pathogenesis. *CD247* encodes the T-cell receptor zeta (CD3 $\zeta$ ) subunit, a component of the T-cell receptor (TCR)-CD3 complex<sup>15</sup>. The CD3 $\zeta$  chain plays an important role in the assembly of the TCR-CD3 complex and its transport to the cell surface and is crucial to receptor signaling function. It has been observed that the expression of the CD3 $\zeta$  chain is altered in chronic autoimmune and inflammatory disorders and that its low expression results in impaired immune response<sup>16–18</sup>. Notably, *CD247* has been associated with susceptibility to systemic lupus erythematosus, another systemic autoimmune disease<sup>19,20</sup>. Moreover, genetic variants in the 3' untranslated region of this gene have shown functional implications leading to a reduced expression of the CD3 $\zeta$  chain that could be manifested in systemic autoimmunity<sup>19</sup>. Therefore, further studies aiming to dissect the exact role of this molecule in SSc will be of interest.

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

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

**URLs.** Illumina iControlDB database, <http://www.illumina.com/science/icontribdb.ilmn/>.

## METHODS

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

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

## ACKNOWLEDGMENTS

This work was supported by the following grants: T.R.D.J.R. was funded by the VIDI laureate from the Dutch Association of Research (NWO) and Dutch Arthritis Foundation (National Reumafonds). J.M. was funded by GEN-FER from the Spanish Society of Rheumatology, SAF2009-11110 from the Spanish Ministry of Science, CTS-4977 from Junta de Andalucía, Spain and in part by Redes Temáticas de Investigación Cooperativa Sanitaria Program, RD08/0075 (RIER) from Instituto de Salud Carlos III (ISCIII), Spain (J.M.). R.B. is supported by the I3P Consejo Superior de Investigaciones Científicas program funded by the 'Fondo Social Europeo'. B.Z.A. is supported by the Netherlands Organization for Health Research and Development (ZonMW grant 016.096.121). B.K. is supported by the Dutch Diabetes Research Foundation (grant 2008.40.001) and the Dutch Arthritis Foundation (Reumafonds, grant NR 09-1-408). Genotyping of the Dutch control samples was sponsored by US National Institutes of Mental Health funding, R01 MH078075 (R.O.A.). The German controls were from the PopGen biobank (to B.K.). The PopGen project received infrastructure support through the German Research Foundation excellence cluster 'Inflammation at Interfaces'. The US analyses were supported by the US National Institutes of Health and National Institute of Arthritis and Musculoskeletal Diseases (NIH-NIAMS) R01 AR055258, Two-Stage Genome Wide Association Study in Systemic Sclerosis, (M.D.M.) and by the NIH-NIAMS Center of Research Translation (CORT) in SSc (P50AR054144) (F.C.A.), the NIH-NIAMS SSc Family Registry and DNA Repository (N01-AR-0-2251) (M.D.M.), University of Texas Health Science Center-Houston Center for Clinical and Translational Sciences (Houston Clinical and Translational Science Awards Program) (NIH-National Center for Research Resources 3UL1RR024148) (F.C.A.), NIH-NIAMS K08 Award (K08AR054404) (S.K.A.), SSc Foundation New Investigator Award (S.K.A.).

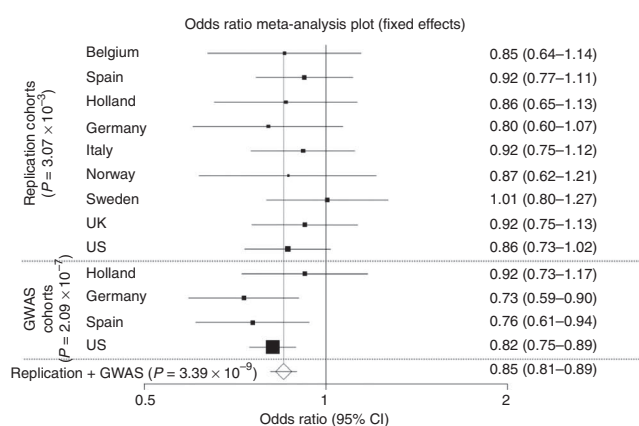
## AUTHOR CONTRIBUTIONS

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

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

**Interpretation and analysis of results:** T.R.D.J.R., O.G., B.R., J.-E.M., B.Z.A., R.P.M., J.Y., Y.H., S.-F.W., R.v.t.S., P.G., A.T.L., C.I.A., S.K.A., B.P.C.K., J.M., M.D.M., A.I., P.C., S.A., P.K.G.

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



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

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

#### COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

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

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

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

## ONLINE METHODS

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

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

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

All cases either met the American College of Rheumatology Preliminary criteria for the classification of SSc or had at least three of the five CREST (calcinosis, Raynaud's phenomenon, esophageal dysmotility, sclerodactyly, telangiectasias) features<sup>22</sup>. Main clinical features of SSc cases are included in **Supplementary Table 5**.

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

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

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

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

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

21. Gourh, P. *et al.* Association of the PTPN22 R620W polymorphism with anti-topoisomerase I- and anticentromere antibody-positive systemic sclerosis. *Arthritis Rheum.* **54**, 3945–3953 (2006).
22. Anonymous. Preliminary criteria for the classification of systemic sclerosis (scleroderma). Subcommittee for scleroderma criteria of the American Rheumatism Association Diagnostic and Therapeutic Criteria Committee. *Arthritis Rheum.* **23**, 581–590 (1980).
23. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
24. Tian, C. *et al.* European population genetic substructure: further definition of ancestry informative markers for distinguishing among diverse European ethnic groups. *Mol. Med.* **15**, 371–383 (2009).