

## ARTICLE OPEN



# Complement component C4 structural variation and quantitative traits contribute to sex-biased vulnerability in systemic sclerosis

Martin Kerick<sup>1,82</sup>✉, Marialbert Acosta-Herrera<sup>1,2,82</sup>✉, Carmen Pilar Simeón-Aznar<sup>3</sup>, José Luis Callejas<sup>4</sup>, Shervin Assassi<sup>5</sup>, International SSc Group\*, Susanna M. Proudman<sup>6</sup>, Mandana Nikpour<sup>7</sup>, Australian Scleroderma Interest Group (ASIG)\*, PRECISEADS Clinical Consortium\*, Nicolas Hunzelmann<sup>8</sup>, Gianluca Moroncini<sup>9</sup>, Jeska K. de Vries-Bouwstra<sup>10</sup>, Gisela Orozco<sup>11,12</sup>, Anne Barton<sup>11,12</sup>, Ariane L. Herrick<sup>13</sup>, Chikashi Terao<sup>14</sup>, Yannick Allanore<sup>15</sup>, Carmen Fonseca<sup>16</sup>, Marta Eugenia Alarcón-Riquelme<sup>17</sup>, Timothy R. D. J. Radstake<sup>18</sup>, Lorenzo Beretta<sup>19</sup>, Christopher P. Denton<sup>16</sup>, Maureen D. Mayes<sup>5</sup> and Javier Martin<sup>1</sup>✉

Copy number (CN) polymorphisms of complement C4 play distinct roles in many conditions, including immune-mediated diseases. We investigated the association of C4 CN with systemic sclerosis (SSc) risk. Imputed total C4, C4A, C4B, and HERV-K CN were analyzed in 26,633 individuals and validated in an independent cohort. Our results showed that higher C4 CN confers protection to SSc, and deviations from CN parity of C4A and C4B augmented risk. The protection contributed per copy of C4A and C4B differed by sex. Stronger protection was afforded by C4A in men and by C4B in women. C4 CN correlated well with its gene expression and serum protein levels, and less C4 was detected for both in SSc patients. Conditioned analysis suggests that C4 genetics strongly contributes to the SSc association within the major histocompatibility complex locus and highlights classical alleles and amino acid variants of HLA-DRB1 and HLA-DPB1 as C4-independent signals.

npj Genomic Medicine (2022)7:57; <https://doi.org/10.1038/s41525-022-00327-8>

## INTRODUCTION

Systemic sclerosis (SSc) is a chronic immune-mediated inflammatory disease (IMID) more frequently observed in women (female/male ratio ~3.8–11.5) that affects the connective tissue and is associated with considerable morbidity and mortality<sup>1–3</sup>. The heterogeneous clinical manifestations of SSc are characterized by functional and structural vasculopathy, fibrosis of the skin and internal organs, in addition to inflammatory and immunological alterations like auto-antibody production<sup>1</sup>. The individual genetic background, together with environmental risk factors and epigenetics factors, play an important role in the pathogenesis of the disease<sup>4,5</sup>.

A recent genome-wide association study (GWAS) has identified new genes and pathways implicated in the development and progression of SSc<sup>6</sup>. Similar to other IMIDs, these genetic variations account for a limited portion of estimated heritability<sup>7</sup>, making clear that additional genetic variants remain to be found with the potential to bring novel insights into disease etiology and

pathogenesis. In this sense, structural variants not captured by GWAS, such as copy number (CN) polymorphisms, which have been implicated in the etiology of several diseases<sup>8,9</sup>, could contribute substantially to the genetic risk of SSc. Several CN variants in immunological genes have been found to be associated with autoimmune diseases<sup>10–13</sup>, although technical limitations and the complexity of CN polymorphisms have reduced the impact of their analysis in understanding autoimmunity<sup>14,15</sup>.

The complement system plays an important role in innate immunity and forms a bridge to the adaptive immune response<sup>16–18</sup>. Functional abnormalities in the complement system have been widely described in rheumatic diseases, such as rheumatoid arthritis (RA) or systemic lupus erythematosus (SLE), and to a lesser extent in SSc<sup>19</sup>. Furthermore, genetic variability in several complement components may contribute to the development of inflammatory and autoimmune diseases<sup>20,21</sup>.

<sup>1</sup>Department of Cell Biology and Immunology, Institute of Parasitology and Biomedicine López-Neyra, CSIC, Granada, Spain. <sup>2</sup>Systemic Autoimmune Disease Unit, Hospital Clínico San Cecilio, Instituto de Investigación Biosanitaria Ibs. GRANADA, Granada, Spain. <sup>3</sup>Department of Internal Medicine, Valle de Hebrón Hospital, Barcelona, Spain. <sup>4</sup>Department of Internal Medicine, Hospital San Cecilio, Granada, Spain. <sup>5</sup>Department of Rheumatology, The University of Texas Health Science Center at Houston, Houston, TX, USA. <sup>6</sup>Rheumatology Unit, Royal Adelaide Hospital and University of Adelaide, Adelaide, SA, Australia. <sup>7</sup>The University of Melbourne at St. Vincent's Hospital, Melbourne, VIC, Australia. <sup>8</sup>Department of Dermatology, University of Cologne, Cologne, Germany. <sup>9</sup>Department of Clinical and Molecular Science, Università Politecnica delle Marche e Ospedali Riuniti, Ancona, Italy. <sup>10</sup>Department of Rheumatology, Leiden University Medical Center, Leiden, The Netherlands. <sup>11</sup>Center for Genetics and Genomics Versus Arthritis, Division of Musculoskeletal and Dermatological Sciences, School of Biological Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, UK. <sup>12</sup>NIHR Manchester Biomedical Research Center, Manchester University NHS Foundation Trust, Manchester, Greater Manchester, UK. <sup>13</sup>Division of Musculoskeletal and Dermatological Sciences, The University of Manchester, Northern care Alliance NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK. <sup>14</sup>Laboratory for Statistical and Translational Genetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa, Japan. <sup>15</sup>Department of Rheumatology A, Hospital Cochin, Paris, Ile-de-France, France. <sup>16</sup>Center for Rheumatology, Royal Free and University College Medical School, London, UK. <sup>17</sup>Center for Genomics and Oncological Research (GENYO), Pfizer-University of Granada-Andalusian Regional Government, Granada, Spain. <sup>18</sup>Department of Rheumatology and Clinical Immunology, University Medical Center Utrecht, Utrecht, The Netherlands. <sup>19</sup>Referral Center for Systemic Autoimmune Diseases, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico di Milano, Milan, Italy. <sup>82</sup>These authors contributed equally: Martin Kerick, Marialbert Acosta-Herrera. \*Lists of authors and their affiliations appear at the end of the paper. ✉email: mkerick@ipb.csic.es; m.acostaherrera@ipb.csic.es; javiermartin@ipb.csic.es

Complement component 4 (*C4*), encoded by two closely linked, highly polymorphic genes *C4A* and *C4B* within the major histocompatibility complex (MHC) class III region on chromosome 6, is an important protein in the classical and lectin complement activation pathways, which are major effectors for controlling microbial infections and for promoting clearance of apoptotic cells and soluble immune complexes<sup>17,19</sup>. *C4A* and *C4B* encode proteins with distinct affinities for their molecular targets<sup>21–23</sup> and present variability in genomic copy number (CN)<sup>24</sup> and length. The long form is determined by the presence of a 6.4-kb human endogenous retrovirus K (HERV-K) element in intron 9 of both genes. *C4* CN studies in autoimmunity have been mainly focused on SLE providing no definitive results<sup>14,22,25</sup> in part due to the complexity of the genetic variation of *C4A* and *C4B* and the high linkage disequilibrium (LD) of the MHC locus, which contains the HLA genes, the strongest genetic associations with most autoimmune diseases including SLE, RA, and SSc<sup>6,26,27</sup>. A recent study, however, managed to attribute most of the genetic association of SLE and Sjögren's syndrome (SjS) of the MHC locus to *C4*<sup>28</sup>. Of note, the same study made the complex genetic variation of *C4A* and *C4B* accessible through the imputation of SNP data using a large multi-ancestry panel of 2530 reference haplotypes<sup>28,29</sup>.

CN variations at *C4* genes have been implicated as a source of sexual dimorphism in two systemic autoimmune diseases, SLE and SjS<sup>28</sup>, diseases that have a higher prevalence in women than men. In light of the above, we set out to investigate the contribution of *C4* CN to SSc using data from the largest GWAS cohort of SSc, combined with genetic, RNA-sequencing, and *C4* serum data from an independent cohort.

## RESULTS

### Experimental design

To investigate the role of *C4* genetics in SSc, we analyzed two independent cohorts of European descent (see “Methods”, Supplementary Fig. 1 and Supplementary Table 1). We used the first cohort (genetic data,  $N = 26,633$ , 34% SSc, 10 countries) to determine the association of *C4* CN to SSc. The second cohort (genetic, whole blood expression, and *C4* serum data,  $N = 857$ , 39% SSc, 9 countries) was analyzed to detect *C4* expression quantitative trait loci (eQTL) and for *C4* expression- and *C4* protein level- modeling. Expression-model eQTLs from the second cohort that explained expression variance were subsequently used in the first cohort as co-factors to determine residual genetic association to SSc unrelated to *C4* genetics in the MHC region on chromosome 6. Finally, we utilized classical HLA alleles and amino acids (AAs) to model remnant *C4*-independent association in the MHC region.

### *C4* haplotype diversity and its correlation with classical HLA alleles

The complex genetic variation of *C4A* and *C4B*, which consists of many haplotypes with different numbers of *C4A* and *C4B* genes was recently made accessible for analysis in large cohorts<sup>28</sup>. We imputed 29 *C4* haplotypes independently in both cohorts. Haplotype frequencies were found to be similar in both cohorts and comparable to published results<sup>28</sup> but varied substantially across countries (Supplementary Fig. 2 and Supplementary Table 2A). We determined nine *C4* haplotypes to have a moderate to strong correlation ( $r^2 > 0.4$ ) to at least one classical HLA allele (Supplementary Dataset 1). Correlation to classical HLA alleles associated to SSc on a genome-wide level was found to be small ( $r^2 < 0.3$ ) for most haplotypes with the exceptions of *HLA-B\*08:01* and *HLA-C\*07:01* which had a strong correlation ( $r^2 = 0.82$  and  $0.67$ , respectively) with the BS haplotype and *HLA-C\*16:01* with AL-BS (H2111B) of 0.63 (Supplementary Dataset 1).

### The association between *C4* copy number variants and SSc is modified by sex and HERV-K

We found a higher *C4* CN to be protective for SSc (Fig. 1a and Table 1). Less than four copies of *C4* were found in 28.6% of SSc patients, and 2.8% had less than three copies. Interestingly, less than four copies of *C4* were found in 28% of female and 32% of male patients. In a simple additive model *C4A*, *C4B* and HERV-K CNs exhibited 1.8-fold variation in their relative risk of SSc (95% confidence interval (95% CI), [1.51–2.33];  $P = 1.04 \times 10^{-16}$ ). Logistic-regression analysis estimated a rather small difference between the protection afforded by each copy of *C4A* (odds ratio (OR) = 0.73; 95% CI = 0.69–0.78) or *C4B* (OR = 0.82; 95% CI = 0.77–0.87) (Table 1). We replicated our calculation in the second cohort and performed a meta-analysis, showing consistent results (Table 1).

The number of subjects in the first cohort permits the simple additive model to be expanded to a more complex one investigating the predictors that influence each other. We found evidence for three two-way interactions: an interaction of *C4A* with *C4B*, a sexual dimorphism of *C4A* and *C4B*, and an interaction of HERV-K CN with *C4A* and *C4B* (Supplementary Table 3). The full model  $\log(\text{risk}) \sim b_1 \text{ Sex} + b_2 \text{ C4A} + b_3 \text{ C4B} + b_4 \text{ HERV-K} + b_5 \text{ Sex:C4A} + b_6 \text{ Sex:C4B} + b_7 \text{ HERV-K:C4A} + b_8 \text{ HERV-K:C4B} + b_9 \text{ C4A:C4B}$  predicted 7.7-fold variation in the relative risk of SSc. The following analyses have been derived from this complex model and its calculated coefficients.

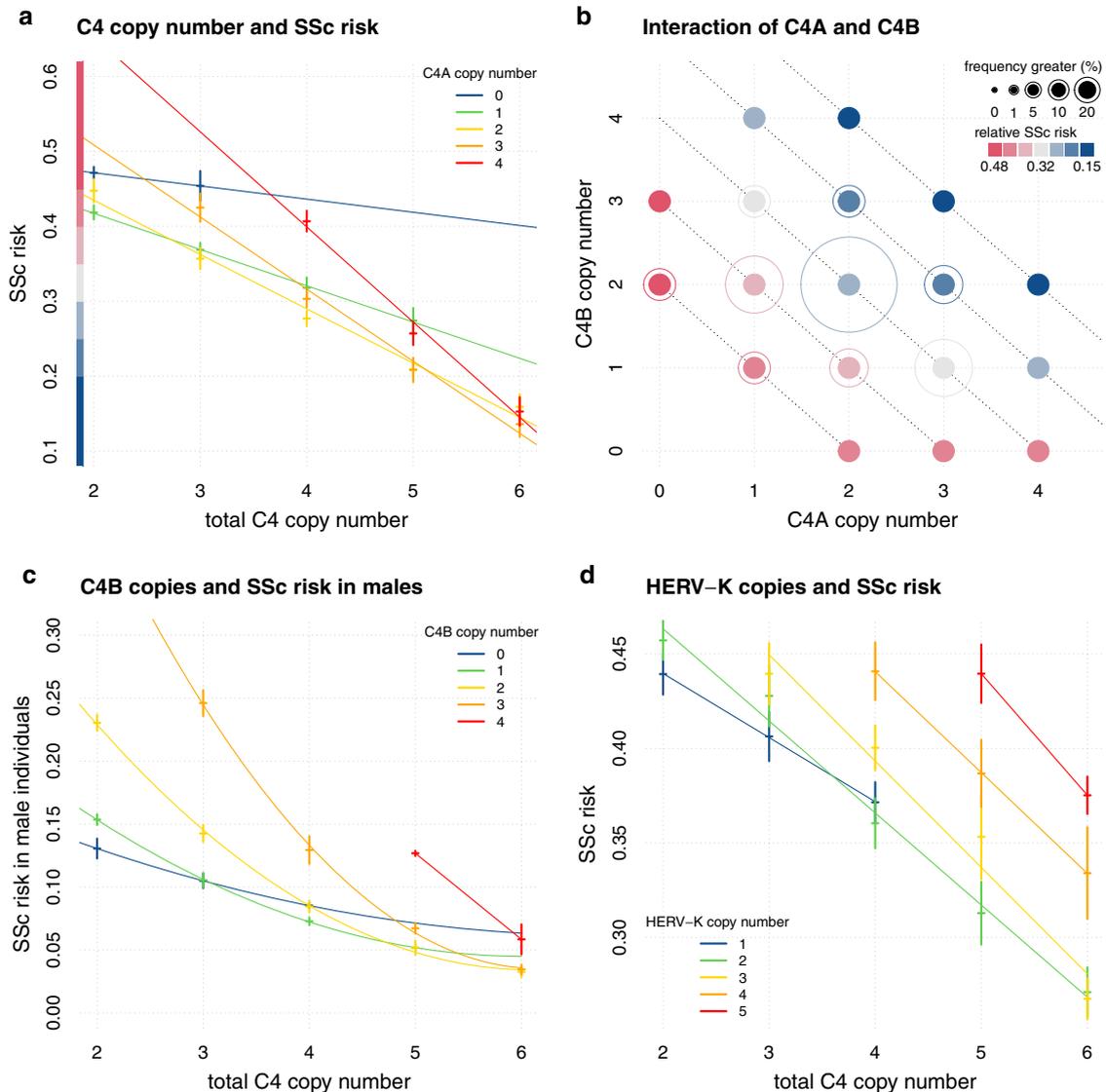
The relationship between SSc risk and *C4A* and *C4B* gene CNs exhibited consistent, logical patterns across the 18 different CN combinations of *C4A* and *C4B* (Fig. 1a, b), which is based on an interaction between *C4A* and *C4B* CNs that was suggested by logistic-regression analysis ( $b_{\text{C4A:C4B}} = -0.14$ ,  $P = 2.1 \times 10^{-5}$ ). While higher total *C4* CN provide protection, strong deviations from the 1:1 ratio of *C4A* and *C4B* are of higher risk like e.g., four copies of *C4A* and zero copies of *C4B* (Fig. 1b).

We found evidence for a sexual dimorphism for *C4A* and *C4B* but not for HERV-K CN. Stratified analysis showed *C4A* to be more protective in men ( $b_{\text{male}} = -0.49$ ,  $P = 1.7 \times 10^{-9}$  vs  $b_{\text{female}} = -0.27$ ,  $P = 6.5 \times 10^{-12}$ ) while *C4B* showed statistical evidence to be protective only in women ( $b_{\text{female}} = -0.22$ ,  $P = 1.1 \times 10^{-10}$  vs  $b_{\text{male}} = -0.09$ ,  $P = 0.23$ ) (Table 1). Indeed, *C4B* CNs of two or higher seem to augment the risk for men (Fig. 1c). Logistic regression with interaction terms confirmed our results for the *C4A*:sex interaction, while the significance of the *C4B*:sex interaction was only suggestive ( $b_{\text{male:C4A}} = -0.17$ ,  $P = 0.0034$ ,  $b_{\text{male:C4B}} = 0.13$ ,  $P = 0.065$ ) (Supplementary Table 3). For total *C4* CN, we did not find a significant sex bias ( $b_{\text{male:C4}} = -0.1$ ,  $P = 0.074$ ). We calculated the statistical power for *C4B* association in males to be 0.41 and for the interaction of *C4B*:sex to be 0.33.

HERV-K copies generally augment the risk for SSc (Fig. 1d and Table 1). The protection afforded by *C4A* and *C4B* CNs are affected by HERV-K copies in a slightly different way. Protection associated with each *C4B* copy is affected more strongly by HERV-K ( $b_{\text{HERV-K:C4B}} = 0.085$ ,  $P = 0.053$ ) in comparison to *C4A* ( $b_{\text{HERV-K:C4A}} = 0.042$ ,  $P = 0.03$ ). A logistic-regression analysis with separate terms for the long and the short forms of *C4A* and *C4B* confirmed that the short forms confer more protection than the long forms (Table 1).

### *C4* copy number affects *C4* expression and *C4* protein levels in whole blood

We investigated the effect of *C4* and HERV-K CN on whole blood expression levels of *C4* ( $C4 = C4A + C4B$ , Fig. 2a). Specific analysis for *C4A* and *C4B* showed significant positive correlation for both *C4A* CN ( $r_{\text{pearson,C4A}} = 0.15$ ,  $P = 1.2 \times 10^{-5}$ ) and *C4B* CN ( $r_{\text{pearson,C4B}} = 0.34$ ,  $P = 1.7 \times 10^{-24}$ ) with *C4A* or *C4B* expression. HERV-K CN had a weakening effect on *C4* expression at all *C4* CN levels (Fig. 2a). *C4A* or *C4B* expression models performed better if they contained separate CNs for long (L) and short (S) versions of *C4*



**Fig. 1 C4 and HERV-K copy numbers and Systemic Sclerosis risk.** **a** depicts relative systemic sclerosis (SSc) risk vs total C4 copy number stratified by C4A CN. The SSc risk score is calculated per individual as the sum of effect sizes (betas) multiplied with the design matrix. Betas of *C4A*, *C4B* and *C4A:C4B* were taken from the most complex model “d” (see “Methods”). Crosses are calculated as average relative risk per rounded C4 CN  $\pm$  2 standard deviations (y axis). Linear regression lines are colored by C4A CN and drawn to visualize the interaction effect of C4A and C4B. The y axis contains a color code to aid a comparison with **(b)**. **b** depicts the relative SSc risk of combinations of C4A and C4B CNs. Relative risk is calculated as in **(a)**. Outer circles are drawn according to population frequency ranges of each C4A, C4B combination and highlight more common combinations. Diagonal dotted lines help to identify combinations of equal total C4 CN. **c** depicts relative SSc risk in male individuals vs total C4 CN stratified by C4B CN. Relative risk is calculated like in **(a)** using effect sizes of *C4A*, *C4B*, *C4A:C4B*, *Sex:C4A*, and *Sex:C4B*. Crosses are calculated as average relative risk per rounded C4 CN  $\pm$  2 standard deviations (y axis). Cubic regression lines are colored by C4B CN and drawn to visualize the interaction effect of C4A and C4B. **d** depicts relative SSc risk vs total C4 CN stratified by HERV-K CN. Relative risk is calculated like in **(a)** using effect sizes of *C4A*, *C4B*, *C4A:C4B*, *HERV-K:C4A*, and *HERV-K:C4B*. Crosses are calculated as average relative risk per rounded C4 CN  $\pm$  2 standard deviations (y axis). Linear regression lines are colored by HERV-K CN.

(AS, AL, BS, BL) instead of C4A plus C4B plus total HERV-K CNs. Interestingly, expression models were best if CNs of C4B were included in the model of C4A and vice versa. About 21% (27%) of expression variance of C4A (C4B) can be attributed to C4A and C4B CNs (Supplementary Table 4B, C).

Using C4A and C4B eQTLs alone, we were able to explain up to 42% (C4B: 38%) of expression variance with 19 (C4B: 15) SNPs (Supplementary Table 4B, C). This seems to be the upper bound of C4 expression variance explained by C4 genetics as C4A and C4B CN and eQTLs together could not explain more than 40% of expression variance albeit with only 12 (C4B: 13) additional SNPs (Supplementary Tables 4B, C). C4 eQTLs seem to integrate C4 and HERV-K CN information. Indeed, copy numbers of *C4A<sub>Short</sub>*, *C4A<sub>Long</sub>*,

*C4B<sub>Short</sub>* and *C4B<sub>Long</sub>* can be predicted well ( $r^2_{C4AL} = 0.5$ ,  $r^2_{C4AS} = 0.58$ ,  $r^2_{C4BL} = 0.54$ ,  $r^2_{C4BS} = 0.67$ , all  $P < 2.2 \times 10^{-16}$ ) using C4 eQTLs forward selected to explain C4A or C4B expression variance.

Blood serum concentrations of C4 protein were well correlated with C4 CNs ( $r_{Pearson} = 0.25$ ,  $P = 1.3 \times 10^{-12}$ , Fig. 2b). Regression analysis determined that C4 and HERV-K CNs, sex, age, and SSc explained about 12% of C4 serum concentration variance ( $P = 1.4 \times 10^{-20}$ ) with HERV-K copies again weakening C4 levels ( $b_{HERV-K} = -0.03$ ,  $P = 9 \times 10^{-7}$ ) (Fig. 2b and Supplementary Table 5). CN-corrected serum C4 levels determined men to have more C4 protein than women ( $b_{male} = 0.04$ ,  $P = 4.6 \times 10^{-4}$ ) independent of disease (Fig. 2e and Supplementary Table 5). Individuals with

**Table 1.** Logistic-regression analysis for total C4, C4A, C4B, and HERV-K copy numbers.

Model	Model terms	1st cohort (N = 26,633)			2nd cohort (N = 857)			Meta-analysis		
		Beta	s.e.	P	Beta	s.e.	P	Beta	s.e.	P
a: all	total C4	−0.23	0.03	6.3E-17	−0.20	0.13	0.12	−0.23	0.03	1.9E-17
	HERV-K	0.12	0.02	2.9E-10	0.16	0.10	0.10	0.12	0.02	7.9E-11
b: all	C4A	−0.31	0.03	3.7E-19	−0.36	0.18	0.04	−0.31	0.03	4.7E-20
	C4B	−0.20	0.03	7.0E-11	−0.14	0.14	0.33	−0.19	0.03	4.6E-11
b: female	HERV-K	0.16	0.02	2.6E-13	0.23	0.11	0.04	0.16	0.02	3.4E-14
	C4A	−0.27	0.04	6.5E-12	−0.81	0.33	0.01	−0.28	0.04	2.6E-12
b: male	C4B	−0.22	0.03	1.1E-10	−0.10	0.27	0.70	−0.22	0.03	2.3E-13
	HERV-K	0.15	0.02	1.4E-10	0.41	0.19	0.03	0.15	0.02	1.5E-14
b: male	C4A	−0.49	0.08	1.7E-09	−1.19	0.48	0.01	−0.51	0.08	1.1E-10
	C4B	−0.09	0.07	2.3E-01	−0.73	0.43	0.09	−0.11	0.07	1.2E-01
b: male	HERV-K	0.18	0.05	2.8E-04	0.66	0.28	0.02	0.19	0.05	7.5E-05
	C4A	−0.29	0.18	1.0E-01	−0.43	0.32	0.18	−0.32	0.16	3.9E-02
c: all	C4ALong	−0.16	0.02	2.0E-10	−0.11	0.18	0.55	−0.16	0.02	1.1E-15
	C4BShort	−0.20	0.03	1.4E-10	−0.12	0.28	0.67	−0.20	0.03	2.5E-11
c: all	C4BLong	−0.04	0.03	1.8E-01	0.09	0.15	0.56	−0.04	0.03	2.3E-01

Depicted are beta values from the logistic-regression analysis of three different models (blocks of rows, see “Methods”). All models contained sex and five genetic principal components as co-variables. Logistic-regression analysis for the first cohort additionally contained cohort as co-variable. Model b was also calculated separately for females and males. Models contain copy numbers as calculated from the imputed C4 alleles per individual as dosages.

SSc had less CN-corrected C4 protein than healthy subjects independent of sex ( $b_{SSc} = -0.02$ ,  $P = 0.012$ ).

Despite that only about a third of SSc patients have less than four copies of C4, we found significant downregulation of C4 expression ( $p_{female} = 0.001$ ,  $p_{male} = 0.003$ ) and C4 protein levels ( $P = 0.004$ ) in SSc patients compared to healthy controls (Fig. 2c, d).

### C4 genetics can explain a part of the SSc association to the MHC region

We performed conditional association analysis for genetic markers across the MHC genomic region. Conditioning on C4A, C4B, and HERV-K CN or on a risk score calculated using the complex C4 CN interaction model derived above, showed an impact on residual association levels limited to the vicinity of the C4 gene (Supplementary Fig. 3B, C). In addition, we calculated a C4 risk score recently proposed for SLE and SjS based only on C4A and C4B CN (as:  $risk = (2.3)C4A\ CN + C4B\ CN$ )<sup>28</sup>. We applied the SLE/SjS score for conditional association analysis in the first dataset and found almost no effect (Supplementary Fig. 3A).

Given the strong association of C4 CN with SSc and its rather local impact in conditional analysis, we focused on C4 eQTLs as potential modifiers of C4 CN risk as both CN and eQTLs affect C4 expression levels. Our analysis above suggests that the predictive power of C4 CN and C4 eQTLs on C4 expression levels is at least partly redundant. In fact, C4 eQTLs alone can explain more C4 expression variance than C4 CNs and C4 eQTLs are significant predictors for copy numbers of C4. We obtained 10,680 eQTLs of C4A and C4B from the GTEx v8 database and found that 37% of these are associated to SSc with  $p_{GWAS} < 10^{-5}$  using the first dataset (Fig. 3a). This encouraged us to find independent C4 eQTLs using forward selection to explain their contribution to the MHC association to SSc.

Conditioned association analysis for genetic markers across the MHC genomic region on ten independent C4 eQTL SNPs rendered most association to SSc nonsignificant ( $P > 5 \times 10^{-8}$ ), except for a peak around HLA-DPB1, the initial association of which was enhanced by conditioning (Fig. 3b). While conditioning on 13 independent C4A exclusive eQTL SNPs had a similar effect on the residual association profile, conditioning on 12 C4B exclusive eQTL

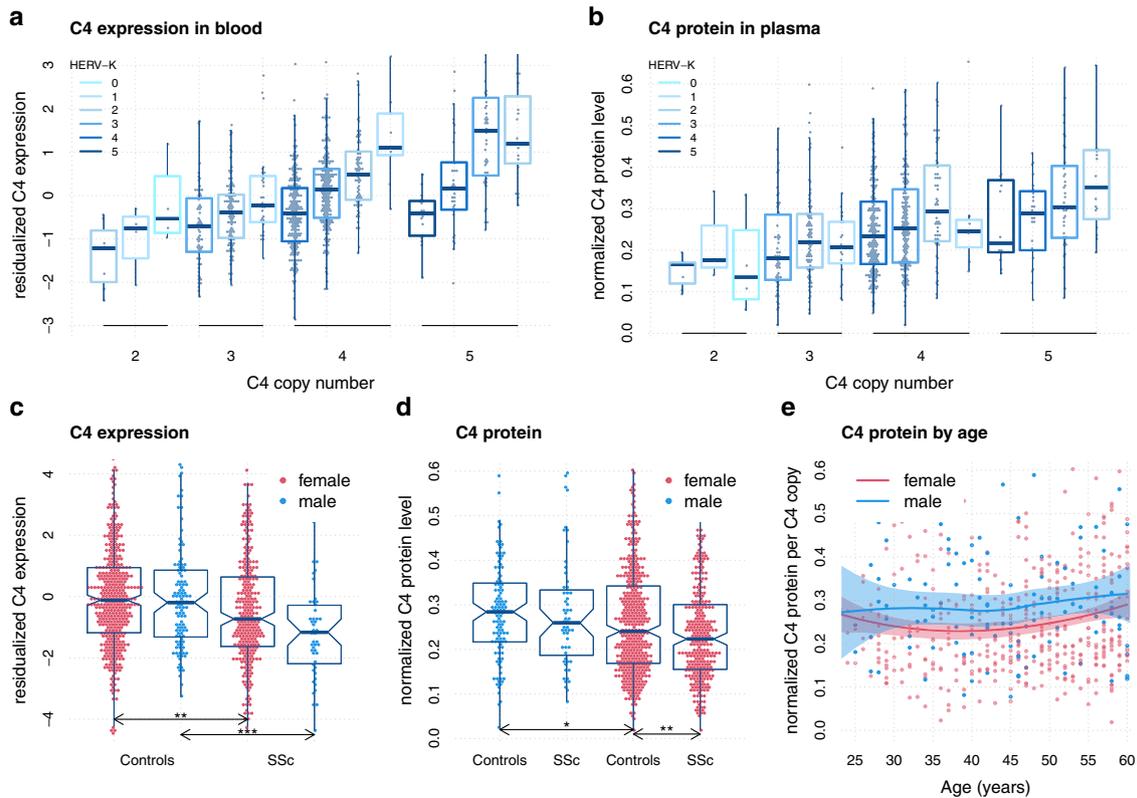
SNPs had a smaller effect with residual association ( $P < 5 \times 10^{-8}$ ) to SSc in three regions centered on HLA-DPB1, HLA-DRB1, and HLA-B (Supplementary Fig. 4B, C), which suggests a stronger contribution of C4A to SSc.

While the ten independent C4 eQTL SNPs absorb as much SSc association as possible by design (forward selection), they might have been selected due to the extensive genetic linkage in the MHC region without being implicated in SSc pathogenesis. We therefore asked if SSc association with the MHC could be explained by C4 eQTLs selected to explain C4A and C4B expression variance in the second dataset. Using expression-model eQTL SNPs, selected in the second cohort to explain expression variance, as co-factors in the first cohort to determine residual genetic association with SSc again rendered most MHC association with SSc nonsignificant ( $P > 5 \times 10^{-8}$ ), except for the peaks shown in Fig. 3c centered on HLA-DPB1 and HLA-DRB1.

### Remaining MHC signal after conditioning on C4 genetics highlights HLA-DRB1 and HLA-DPB1

Having attributed most SSc association within the MHC to C4 genetics, we investigated which classical HLA alleles and HLA amino acids (AA) demonstrated C4-independent association to SSc. Conditioning on expression derived independent C4 eQTLs results in residual significance ( $P < 5 \times 10^{-8}$ ) for classical alleles and AAs of HLA-DRB1, HLA-DPB1, HLA-DPA1 and HLA-DQA1, HLA-DQB1, and HLA-B (Supplementary Dataset 2A, B). Forward selection to derive independent residual signals marked HLA classical alleles for HLA-DRB1 (\*07:01, \*11:03, \*11:04, \*13:01), HLA-DPB1 (\*13:01, \*26:01, \*40:01, \*06:01), HLA-DQA1 (\*04:01), and HLA-DQB1\*05:01.

We tested if classical alleles and AAs of HLA-DRB1 and HLA-DPB1 alone could explain residual C4-independent association. We found that 9 HLA classical alleles (HLA-DRB1: \*11:04, \*08:01, \*07:01, \*13:01, \*11:03 and HLA-DPB1: \*13:01, \*26:01, \*40:01, \*06:01) together with 16 expression-model-derived independent C4 eQTLs can explain almost all associations of SSc with the MHC region (Supplementary Fig. 5B). The same is true for AAs of HLA-DRB1 and HLA-DPB1, which can also explain residual C4-independent association with eight independent AAs (HLA-



**Fig. 2** **C4 expression and C4 protein concentrations in whole blood.** **a** depicts residualized total C4 expression levels by total C4 copy number (CN) stratified by HERV-K CN. C4 expression is calculated as the sum of C4A and C4B expression as obtained by RNA-Sequencing. The residualized expression has been calculated by regressing out 20 (18) principal components for controls and cases, respectively. Data has been grouped by rounded C4 and HERV-K CN dosage. **b** depicts normalized C4 protein levels in plasma by total C4 CN stratified by HERV-K CN. C4 protein levels have been normalized across 10+ laboratory sites. **c** depicts residualized total C4 expression levels (like in **a**) for SSc and controls, stratified by sex. Significant comparisons are highlighted by asterisk (\* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ ). **d** depicts normalized C4 protein levels (like in **b**) for SSc and controls, stratified by sex. Significant comparisons are highlighted by asterisk (\* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ ). **e** depicts normalized C4 protein levels in blood from 119 adult men (blue) and 447 adult women (red) as a function of age with locally estimated scatterplot smoothing (LOESS). Protein levels are normalized to the number of C4 gene copies in an individual's genome. All boxplot are drawn with default settings in R 4.0.3: lines are defined as first, second and third quartile (Q1, Q2, Q3), whiskers depict the most extreme data points within Q1–1.5 interquartile range (IQR), and Q3 + 1.5 IQR. Boxplot notches are defined as 95% confidence interval of the median.

DRB1: 375Y, 58A, 74RL, 96E, and HLA-DPB1: 8, 76I, 91H, 96X) (Fig. 3d).

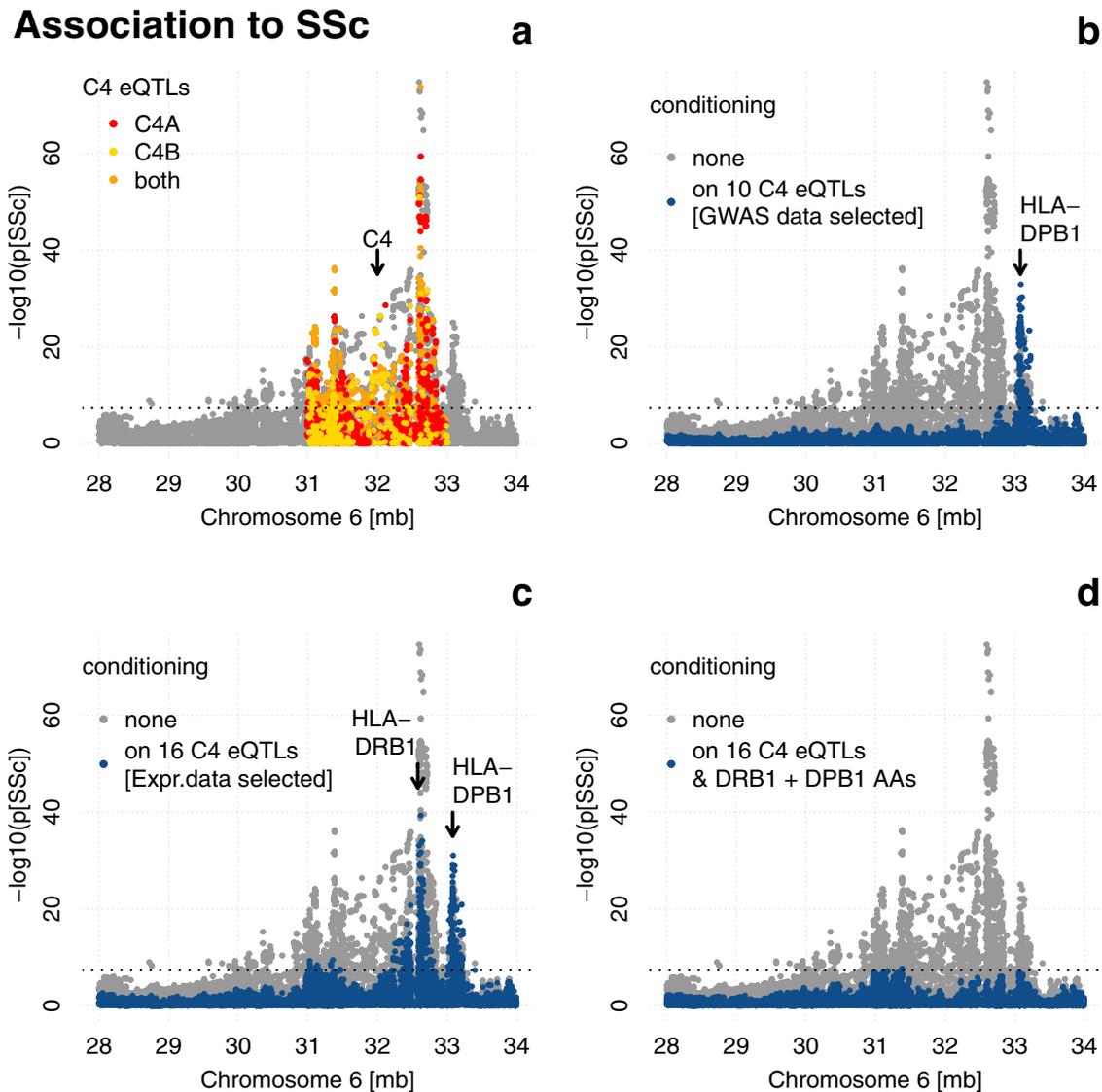
To complement our analysis, we repeated our search for residual C4-independent association to SSc, this time conditioning on the C4 genetic signal which was not derived from the expression dataset but from the first dataset as ten independent C4 eQTL signals as described. Repeating the forward selection of AAs or classical HLA alleles conditioning on the ten first dataset-derived independent C4 eQTLs signals resulted in four independent AAs from HLA-DPB1 (9F, 76I, 91H, 96X) or five independent alleles (HLA-DRB1: \*11:04 and HLA-DBP1: \*13:01, \*26:01, \*28:01, \*30:01) (Supplementary Fig. 4D, E) supporting the role of HLA-DRB1 and HLA-DPB1 as a C4-independent association.

## DISCUSSION

In this study, we found a strong association of low C4 and high HERV-K CN with SSc in two independent SSc datasets and their meta-analysis, supporting the protective role of C4 copies in IMIDs. C4A gene copies were slightly more protective than C4B as has been shown in SLE and SjS<sup>28</sup> but our data suggest a complex interaction of C4A and C4B CNs that has to be evaluated in the context of HERV-K copies and sex. We found that in SSc, an equal number of C4A and C4B gene copies grants more protection than (strongly) imbalanced numbers which we found to be a risk for

SSc (Fig. 1b). Our results might differ from recent observations in SLE and SjS where C4A and C4B copies have been described to act in an additive way, but the authors did not describe the interaction with HERV-K copies in detail<sup>28</sup>.

Our results showed a sexual dimorphism with respect to the protection afforded by C4A and C4B. While in female individuals, C4A copies grant slightly more protection than C4B copies, our data suggest that in male individuals only C4A confers protection while we did not observe a strong effect for C4B. In male individuals, C4B might therefore function like a null allele with respect to protection from SSc as higher CNs of C4B are associated to higher SSc risk (Fig. 1c). However, as the power of our study to detect significant C4B signals in males was limited, and the sex:C4B interaction was only of suggestive significance, we cannot rule out that C4B has a protective effect in males. While C4 alleles have been described to act more strongly in men, no distinction was made between C4A and C4B activity in SLE or SjS in a recent study of similar size<sup>28</sup>. It has been described that activated C4A bonds preferably with protein antigens, such as immune complexes, while activated C4B reacts rapidly with carbohydrate antigens, such as bacterial cell walls<sup>30</sup>. This could partly explain the greater susceptibility to and severity of infections reported in men and the higher incidence of autoimmune diseases in women<sup>31,32</sup>. In addition, low C4B CNs have been associated to cardiovascular disease<sup>33</sup> where the incidence in women is usually



**Fig. 3 MHC region conditional association with systemic sclerosis.** Association is calculated in the first dataset ( $N = 26,633$ ) using logistic regression with cohort, genetic background (PC1-5), and sex as covariates and depicted as position (GRCh38) by significance (Manhattan plot) in gray if no additional covariates were used. The dotted line represents the genome-wide significance cutoff  $P = 5 \times 10^{-8}$ . **a** Manhattan plot with marked *C4* eQTLs obtained from the GTEx v8 database. **b** Manhattan plot with additional conditioning on ten independent *C4* eQTLs, obtained by forward selection in the first dataset, depicted in blue. The arrow marks the position of *HLA-DPB1*. **c** Manhattan plot with additional conditioning on 16 independent *C4* eQTLs (obtained by forward selection to explain expression variance in the second dataset ( $N = 857$ )) depicted in blue. The arrows mark the positions of *HLA-DPB1* and *HLA-DRB1*. **d** Manhattan plot with additional conditioning on 16 independent *C4* eQTLs (obtained by forward selection to explain expression variance in the second dataset) and 8 independent amino acids of *DRB1* and *DPB1* (obtained by forward selection in the first dataset conditioning on 16 independent *C4* eQTLs) depicted in blue.

lower than in men<sup>34</sup>. Interestingly, in a recent paper studying the female-biased expression in human skin, several genes from the complement activation pathway were identified as a molecular signature and in genome-wide co-expression networks<sup>35</sup>. Sexual dimorphism has been extensively reported in vascular physiology and pathophysiology<sup>36</sup>, where women more commonly develop microvascular dysfunction, and in autoimmune-related interstitial lung diseases, where young women are most commonly affected<sup>37</sup>. All of these clinical manifestations, for which the role of *C4* is yet to be elucidated, are hallmarks of SSc.

Our data confirm that *C4* and *HERV-K* CNs are strong predictors of *C4* expression levels in blood and other tissues<sup>29</sup>. While the major site of *C4* expression is the liver<sup>38</sup>, it has been shown that whole blood can be used with some caution as a surrogate tissue for quantitative trait analysis<sup>39-41</sup>. In addition, local complement

production by bone-marrow-derived monocytes and macrophages can restore humoral response in *c4* deficient mice<sup>42</sup>. Interestingly *C4A* and *C4B* expression models both profit strongly from the other gene's CN as a predictor, which supports the genetic interaction between them suggested in this study. Furthermore, the distinction between the long and the short forms of *C4*: AS, AL, BS, and BL as expression predictors instead of *C4A*, *C4B*, and *HERV-K* CNs alone, greatly favors the accuracy of the expression model. This suggests that *HERV-K* acts specifically on the gene where it is located, to suppress its expression, which is supported by studies in brain and serum<sup>22,29,43</sup>. *C4A* and *C4B* CNs were able to explain about 20% of *C4A* and *C4B* expression variance, which is clearly lower than the ability of *C4* eQTLs, which could explain ~40%. Although we most likely over-fitted the expression data, SNPs seem to be the superior instrument in

predicting *C4* expression as they can integrate CN as well as classical eQTL signals. This finding might help to bring *C4* genetics to the clinic in the form of simple genetic tests. The interconnectiveness of *C4* CN and eQTLs is further supported by *C4* eQTLs being able to predict *C4* CNs (AS, AL, BS, BL) with coefficients of determination of  $r^2 > 0.5$ .

*C4* and HERV-K CNs are strong predictors of *C4* protein levels in blood serum<sup>43</sup>. Interestingly, it was reported that men had on average 27% more *C4* protein per *C4* CN than women and that this bias is stronger during reproductive years<sup>28</sup>. While we found the difference between men and women to be smaller; men had on average 17% more *C4* protein per *C4* CN than women; we think our dataset confirms this finding and its timeframe (Fig. 2f) reinforcing the role of *C4* in the differential susceptibility between men and women observed in SSc. The deficiency of *C4* may trigger an inappropriate clearance of apoptotic debris and stimulate chronic activation of myeloid cells. Also, it results in a defect to eliminate autoreactive B-cell clones, and a higher tendency to form self-reactive germinal centers<sup>23</sup> and has been previously associated with more severe SLE with earlier disease onset<sup>44</sup>. We also observed that patients with SSc have lower *C4* serum levels than unaffected individuals even after correcting for *C4* gene CN, suggesting that hypocomplementemia in SSc is not simply due to *C4* genetics but also reflects disease effects on background complement levels<sup>45</sup>.

*C4* expression was clearly down-regulated in SSc patients compared to healthy individuals, as were *C4* protein levels, although to a lesser extent (Fig. 2c, d). This might be explained by the difficulty in standardizing *C4* protein assays across 10+ laboratory sites but might also point to differential mRNA stability adding another layer of complexity yet to be analyzed. Indeed, while we observed clear disease-independent differences of *C4* protein levels between men and women, there was no significant differential expression of *C4* between healthy men and women, which suggests post-translational effects to play a role. In this line it has been proposed that IFN- $\gamma$  increases the stability of *C3* and *C4* mRNA<sup>46</sup> and a recent expression analysis in SSc detected a strong *IFN* signature in a subset of patients<sup>47</sup>.

More than a third of more than 10,000 *C4* eQTLs from the GTEx v8 database are associated with SSc ( $p_{GWAS} < 10^{-5}$ ) and *C4* eQTLs alone can be used to explain most of the association of SSc within the MHC region, further supporting their importance in SSc. Interestingly, *C4A*-specific eQTLs can explain more SSc association than *C4B*-specific eQTLs (Supplementary Fig. 4B, C), which supports a stronger role for *C4A* in SSc. While *C4* eQTLs could in principle be associated with SSc by the strong linkage structure present in the MHC, our data suggest that *C4* eQTLs, forward selected to explain expression variance in blood, can also explain most genetic association with the MHC. Both analyses raise the possibility that *C4* genetics is indeed the main signal on chromosome 6 for SSc, as has been suggested for SLE and SjS<sup>28</sup>, both rheumatic diseases that can co-occur with SSc.

*C4*-independent genetic association with SSc centers on two peaks (Fig. 3c), most of which can be explained by four AAs each of *HLA-DPB1* and *HLA-DRB1* (Fig. 3d). Interestingly, the AA positions for *HLA-DPB1* and *HLA-DRB1* overlap and all 8 AA positions can be associated with four of five binding pockets described for class II HLA molecules<sup>48</sup> likely interfering with (auto-) antigen binding. In addition, three of the *HLA-DRB1* AA positions (37, 58, and 74) are close to sites (30, 60, and 74) which have been described to play a role in binding the consensus antigenic peptide of the topoisomerase I epitope, auto-antibodies to which define the ATA<sup>+</sup> subgroup of SSc patients<sup>49</sup>. Furthermore, we found that the *C4*-independent genetic association with SSc can be explained by 10 independent classical HLA-Alleles instead of AAs, seven of which overlap with a model of nine independent HLA-Alleles recently described<sup>50</sup>, which supports the independence of *C4* and HLA associations with SSc.

Our study has some limitations. First, the number of samples in the second dataset is very low in terms of GWAS. While we were able to replicate the association of *C4A*, *C4B*, and HERV-K CNs, replication of the most complex model was out of reach and needs to be the subject of further study. Second, we did not stratify the patients by clinical or serological subgroups. While our results on HLA-DRB1 AAs, being associated with SSc independently of *C4* genetics, point towards anti-topoisomerase auto-antibodies and probably diffuse cutaneous SSc, the topic is too vast to explore in this study. Third, unfortunately, we could not distinguish *C4A* and *C4B* protein levels in serum, which would have been very useful, to further investigate the sexual dimorphism described. Fourth, the exact amino acid positions and classical alleles from the models calculated in our study might change in the future. New imputation reference panels might provide new associations that could influence the models as forward selection is sensitive to statistical fluctuations. Last, *C4* forms a genetic module termed RCCX with three genes: serine/threonine nuclear protein kinase *RP*, steroid 21-hydroxylase *CYP21*, and extracellular matrix protein tenascin *TNX*<sup>51</sup>. Although we only assessed *C4* CNs associated with SSc, we cannot discard the possibility that this module plays a role in disease susceptibility. Specifically, *TNX* is involved in the maintenance of collagen networks and tissue integrity<sup>52</sup>, as well as in TGF $\beta$  activation and signaling, typical for fibrotic conditions such as SSc<sup>53,54</sup>.

Many rheumatic diseases, including SSc, could benefit from therapies directed toward the complement system. These are currently under active development and are not only focused on inhibitory mechanisms, but on activators or downstream activation fragments<sup>23</sup>. The inhibition of the complement pathway has proven challenging. Eculizumab, a C5 inhibitor, is a complement-targeting approved drug for a variety of vascular disorders and has recently been approved in kidney diseases<sup>55</sup>. Moreover, it has been studied in idiopathic inflammatory myopathies and SSc renal crisis, with promising results<sup>56,57</sup>. Our data suggest that *C4* genetics in SSc, by affecting expression and *C4* protein levels, plays an important role in mediating the genetic association in the MHC locus and might also be involved in the epidemiological sex bias of SSc. This highlights the contribution of the complement to the development of SSc and to autoimmune disorders in general, which could benefit from therapies directed towards the complement system. Our findings might help to bring *C4* genetics to the clinic in the form of simple genetic tests.

## METHODS

### Patients

All patients fulfilled the classification criteria of the 2013 American College of Rheumatology (ACR) or The European League Against Rheumatism (EULAR) or the criteria proposed by LeRoy and Medsger for early SSc<sup>58,59</sup>. CSIC's Ethics Committee approved the study and written informed consent was obtained in accordance with the Declaration of Helsinki.

### Cohorts and datasets

We (re-)analyzed two independent cohorts of European descent:

First cohort: genome-wide genotyped data from 14 independent epidemiological cohorts comprising a total of 28,179 unrelated individuals (9846 SSc patients and 18,333 healthy subjects) from 10 European countries<sup>6</sup>. To identify ancestry outliers ~100,000 quality-filtered independent SNPs were selected from each case-control GWAS cohort. Principal component (PC) analysis was performed using PLINK v1.07. Samples showing >4 standard deviations from the cluster centroids of each cohort were considered outliers and removed from further analyses. The presence of relatives and/or duplicates was assessed by computing identity-by-descent (IBD) estimation using PLINK v1.07. An

individual from each pair of relatives ( $Pi\_Hat > 0.45$ ) or duplicates ( $Pi\_Hat > 0.99$ ) was removed. After exclusion of non-European samples, we recalculated genetic PCs using the merge of all imputed datasets, selecting ~100,000 independent markers using PLINK v1.9. Missing data values due to the different platforms used for genotyping were corrected by PLINK v1.9 (parameter `-correct_for_missingness`). We obtained informative principal components as the visualization of the first two PCs can be interpreted as a “map” of the European continent (Supplementary Fig. 1). Second cohort: this cohort included genome-wide genotyped data, whole blood expression data and blood serum C4 protein concentrations from 333 SSc patients and 524 healthy individuals from 9 European countries<sup>39</sup>. This second cohort is a subset of a larger cohort of seven immune-mediated diseases plus controls described here<sup>60</sup>. Individuals were excluded on the basis of incorrect sex assignment, high missingness (>10%), non-European ancestry (<55% using Frappe<sup>61</sup> and high relatedness (PLINK v1.9  $Pi\_Hat > 0.5$ ). In addition, population stratification was also analyzed by PC analysis selecting ~100,000 independent markers using PLINK v1.9. We obtained informative principal components as the visualization of the first two PCs can be interpreted as a “map” of the European continent (Supplementary Fig. 1).

Basic clinical epidemiological information by cohort can be found in Supplementary Table 1.

### Expression data

Whole blood expression data was obtained from alpha and beta globin depleted (globinCLEAR, Ambion) total RNA. Single end 50 bp stranded sequencing was performed on a HiSeq2500 Illumina within the PRECISEADS consortium<sup>60</sup> and processed with bcl2fastq (Illumina), Cutadapt<sup>62</sup>, STAR v2.5.2 (2-pass default mapping to GRCh19<sup>63</sup>, and RSEM v1.2.31<sup>64</sup> to obtain estimated counts per gene. Raw count data were normalized for quantitative trait analyzes<sup>39</sup>. Briefly: Three genetic principal components (PCs) were regressed out from VSN-normalized<sup>65</sup> raw read count data. Potential non-genetic influences were regressed out for SSc and controls separately by 20 (SSc:18) PCs calculated from inter-sample expression correlation matrices.

### C4 protein data

Human complement C4 serum data was obtained from the PRECISEADS consortium<sup>60</sup> from a turbidimetric immunoassay method according to the manufacturer's recommendations (SPAPLUS analyzer)<sup>66</sup>. A corrective factor was calculated in order to normalize the data between the centers as described<sup>60</sup>.

### Imputation

**SNPs.** For both cohorts, we imputed SNPs from chromosome 6 using the TOPMed reference panel with default settings at <https://imputation.biodatacatalyst.nhlbi.nih.gov/><sup>67</sup>. Stringent QC measures were applied to both cohort's pre-imputation as follows: SNPs with call rates < 0.98; minor allele frequencies (MAFs) < 0.01; and those that deviated from Hardy–Weinberg equilibrium (HWE;  $P < 0.001$  in both case and control subjects) were filtered out from further analysis; samples with call rates < 0.95 were removed. Relatives and/or duplicated samples were removed. Post-imputation quality control included filtered for imputation quality ( $r^2 > 0.3$ ),  $MAF > 0.05$ , and HWE, which<sup>6,39</sup> resulted in 9,068 SSc patients and 17,565 healthy individuals for C4 haplotype imputation for the first cohort.

**C4 haplotypes.** A set of 7021 SNPs TOPMed imputed SNPs were selected as they were (a) imputed in all individuals in both cohorts and (b) overlapped the C4 CN reference panel. C4 haplotype imputation was carried separately for both cohorts using the

software impute4<sup>29</sup> and <https://github.com/freeseek/impute4> and the reference panel downloaded from the dbGaP study accession: phs001992.v1.p1<sup>28</sup>. Weighted imputation accuracy was calculated by multiplying  $r^2_{Allele}$  by Allele frequency in Supplementary Table 2B.

**C4 copy numbers.** Each C4 haplotype carries a specific number of C4 isotypes (C4A, C4B) and HERV-K elements (Supplementary Table 2C). We calculated total C4, C4A, C4B, and HERV-K CN dosages by multiplying the allele dosages of the structural haplotype by the number of copies of each C4 isotype and HERV-K that the haplotype contains. For instance, the haplotype AL-BL contains one C4A gene and one C4B gene and two HERV-K copies. The numbers of short and long forms of C4A and C4B (AL, AS, BL, BS) per haplotype are self-evident for 17 of 29 imputed haplotypes. For the remaining, long and short forms were inferred by the consensus that ~95% of C4A is present in the long form<sup>43,68–70</sup>. The haplotype AL-BS for instance can be coded as 0.95 AL, 0.05 AS, 0.05 BL, and 0.95 BS. CNs per haplotype can be found in Supplementary Table 2C.

**Classical HLA alleles and HLA amino acids (AA).** Data for the classical HLA alleles and AA variants were obtained from the first cohort by imputation using SNP2HLA<sup>71</sup> and the reference panel from the Type 1 Diabetes Genetic Consortium<sup>72</sup>, described in ref.<sup>50</sup>. After genotyping QC, all variants were imputed for each case–control dataset separately in the extended MHC region in chromosome 6. Imputed data were also filtered for 95% success call rate for alleles and amino acids, deviation from HWE considering a  $P$  value of < 0.001 for SNPs in controls and 95% total call rate for individuals<sup>50</sup>.

### Pearson correlation of C4 haplotypes and classical HLA Alleles

Was calculated among the C4 haplotype dosages and the allele dosages from the HLA imputation.

### C4 copy number association analysis

Logistic-regression models from simple to complex were calculated (using the function `glm` in R 4.0.3) to assess the association of total C4 dosage and its isotype dosages with the disease. We included cohort, five genome-wide principal components (PCs) and sex as covariates, assuming their effects were not collinear:

- (a)  $SSc \sim C4 + HERV-K + PC1-5 + cohort + sex$
- (b)  $SSc \sim C4A + C4B + HERV-K + PC1-5 + cohort + sex$
- (c)  $SSc \sim C4A_{short} + C4A_{long} + C4B_{short} + C4B_{long} + PC1-5 + cohort + sex$

The number of subjects in our first cohort permits us to expand the simple additive model to a more complex one investigating the predictors that influence each other. We included three two-way interaction terms in the logistic-regression model:

- (d)  $SSc \sim C4A + C4B + HERV-K + PC1-5 + cohort + sex + C4A:C4B + C4A:HERV-K + C4B:HERV-K + C4A:sex + C4B:sex$

Meta-analysis was conducted with Metasoft<sup>73</sup> using data from model a, b, and c from both datasets.

### Power calculation

Power calculations in CNV studies are problematic because effect sizes and models of the association are based on approximations that may be unrealistic<sup>74</sup>.

C4. Power calculations for C4B in males was carried out using the GAS Power Calculator [\[https://csg.sph.umich.edu/abecasis/](https://csg.sph.umich.edu/abecasis/)

[gas\\_power\\_calculator/](#)]. Here, we calculated the disease allele frequency as  $(\text{sum}(\text{CN } C4B < 2)/N) = 0.26$  and the genotype relative risk as  $e^{0.09} = 1.09$ . Using an additive disease model and 1278 male cases and 6875 male controls, results in a power of 0.406 to detect an association with  $P < 0.05$ .

**C4:sex interaction.** We calculated the power to detect C4B:sex interactions using “powerGWASinteraction”<sup>75</sup> in R 4.0.3 which can treat sex as environmental variable. We used: prevalence = 0.00034, pEnv = 0.306, betaC4B = -0.04, beta.sex = -1.35, beta.c4b:sex = 0.13, caseControlRatio = 0.34, ORgeneEnvironment = 1.03, alpha = 0.05 and alpha1 = 1. pGene probability was calculated as  $(\text{sum}(\text{CN } C4B > 2)/N) = 0.73$ . This results in a power to detect a C4B:sex interaction with  $P$  values  $< 0.05$  of 0.33.

### Calculation of composite C4 risk score for SSc

For each individual (i) a composite C4 risk score can be calculated as the sum of betas “ $S_{b,i}$ ” from the (model-specific) effect sizes multiplied by the design matrix (of CN dosages, sex, interactions.. etc.) of the predictors. An individual relative risk score was then calculated as  $\text{risk}_i = e^{S_{b,i}} / (1 + e^{S_{b,i}})$ .

To visualize the interaction effect of C4A and C4B CNs on relative risk, one multiplies the effect sizes (betas) of the most complex model “d” of C4A, C4B, and C4A:C4B with the design matrix and calculates the relative risk score as above. To visualize the effect of HERV-K on risk we calculated a composite score with model “d” betas and C4A, C4B, HERV-K, C4A:HERV-K, C4B:HERV-K, and C4A:C4B. To visualize the effect of C4B CN in males subjects, we used effect sizes from model “d” for sex, C4A, C4B, sex:C4A, sex:C4B, and C4A:C4B.

### Pearson correlation of C4 CN and C4 expression and C4 serum levels

Was calculated with C4 CN dosages, the PC residualized expression data (see above) and the center corrected C4 protein serum concentrations. For visualization, CN dosages were rounded to integers.

### C4 expression modeling

Total C4 expression was calculated as the sum of C4A and C4B expression. We used the linear model function “lm” in R 4.0.3 to calculate the adjusted coefficient of determination ( $r^2$ ) for each model with C4 CN, C4 CN + C4 eQTLs and C4 eQTLs alone as predictors. Model evolution is noted in Supplementary Table 4A–C. To add C4 eQTLs to the C4 CN model as expression modifiers, we used forward selection. In a stepwise manner, we selected the SNP to add to the model which had the most significant  $P$  value conditioning on all predictors already in the model until no one more SNP was found with  $P < 0.01$ . In the same way, SNPs were selected for the eQTLs only model until no more SNPs were found with  $P < 0.01$ . To select SNPs in the expression (=second) dataset which were to be used for conditioned analysis of the MHC SNPs in the first dataset, forward selection was applied with SNPs which had  $p_{GWAS} < 10^{-5}$  until no more SNP was found with  $P < 0.01$ .

### Modeling of C4 copy numbers using eQTLs

We tested if the eQTLs found to explain C4A or C4B expression variance (see C4 expression modeling) can predict copy number dosages of the long and short forms of C4A and C4B: AS, AL, BS and BL. We used the linear model function “lm” in R 4.0.3 to calculate the adjusted coefficient of determination ( $r^2$ ) for each model with either C4A eQTLs or C4B eQTLs as predictors.

### C4 gene expression analysis in whole blood

Using raw count data, we included disease, blood cell composition, and effective library size (calculated by EdgeR in R 4.0.3) in the final model. While cell type-specific expression changes between SSc and controls were found significant at a nominal level for most cell types, the direction of expression change coincided for all cell types. We decided to report only whole blood expression changes controlling for blood cell composition.

### C4 protein blood serum analysis

We included disease, sex, age, AS, AL, BS, and BL in the final model. The significance for the difference between SSc and controls in men and women was calculated with both the Mann–Whitney test and a  $t$  test.

### Residual association of genetic variants across the MHC region to SSc

We performed conditional association analysis for genetic markers across the MHC genomic region. The first dataset was analyzed. In all models, we included cohort, five genome-wide PCs and sex as basic covariates. Association analysis of MHC region variants was conditioned on the basic covariates plus:

- (1) nothing;
- (2) a risk score:  $2.3 \times C4A \text{ CN} + C4B \text{ CN}$  as proposed;<sup>28</sup>
- (3) covariates from model “b”: C4A CN + C4B CN + HERV-K CN;
- (4) covariates from the most complex model “d” described above;
- (5) C4 (C4A or C4B or both) eQTLs from GTEx v8 (obtained by forward selection in the first dataset until no SNP had  $p_{SNP} < 10^{-5}$ , see Supplementary Tables 4B, 6 and 4C, 6);
- (6) C4A -specific eQTLs from GTEx v8 (obtained by forward selection in the first dataset until no SNP had  $p_{SNP} < 10^{-5}$ ). EQTLs were called C4A-specific if no C4B eQTL was reported in GTEx v8 with  $P < 0.01$  for each SNP;
- (7) C4B-specific eQTLs from GTEx v8 (obtained by forward selection in the first dataset until no SNP had  $p_{SNP} < 10^{-5}$ );
- (8) expression-model SNPs (with  $p_{GWAS} < 10^{-5}$ ) (obtained by forward selection in the second dataset as described above, see Supplementary Tables 4B, 7 and 4C, 7).

### Residual, C4-independent, the association of the MHC region with SSc

After accounting for the contribution of C4 genetics with models “5” or “8” above, we sought to model residual, C4-independent, association of MHC SNPs with (a) forward selection of classical HLA alleles; (b) forward selection of classical HLA alleles of HLA-DRB1 and HLA-DPB1; (c) forward selection of AAs of HLA genes; (d) forward selection of AAs of HLA-DRB1 and HLA-DPB1. Forward selection was carried out until no more HLA allele or AA was found with  $P < 5 \times 10^{-8}$ .

### Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

### DATA AVAILABILITY

Summary statistics of the SSc meta-GWAS are available through the NHGRI-EBI GWAS Catalog (<https://www.ebi.ac.uk/gwas/downloads/summary-statistics>): GCST009131. Data from the PRECISEADS consortium are hosted by ELIXIR Luxembourg <https://elixir-luxembourg.org/> and are available upon request. The access procedure is described on the data landing page (<https://doi.org/10.17881/th9v-xt85>). All other data are contained in the article file and its Supplementary Information.

## CODE AVAILABILITY

All analysis has been performed with either the software described in "Methods" or within R 4.0.3. For logistic-regression analysis, the glm function of R was used.

Received: 3 March 2022; Accepted: 22 September 2022;  
Published online: 05 October 2022

## REFERENCES

- Denton, C. P. et al. Systemic sclerosis. *Lancet* **390**, 1685–1699 (2017).
- Elhai, M. et al. Mapping and predicting mortality from systemic sclerosis. *Ann. Rheum. Dis.* **76**, 1897–1905 (2017).
- Bergamasco, A. et al. Epidemiology of systemic sclerosis and systemic sclerosis-associated interstitial lung disease. *Clin. Epidemiol.* **11**, 257–273 (2019).
- Angioliilli, C. et al. New insights into the genetics and epigenetics of systemic sclerosis. *Nat. Rev. Rheumatol.* **14**, 657–673 (2018).
- Truchetet, M. E. et al. Current concepts on the pathogenesis of systemic sclerosis. *Clin. Rev. Allergy Immunol.* <https://doi.org/10.1007/s12016-021-08889-8> (2021).
- Lopez-Isac, E. et al. GWAS for systemic sclerosis identifies multiple risk loci and highlights fibrotic and vasculopathy pathways. *Nat. Commun.* **10**, 4955 (2019).
- Bossini-Castillo, L. et al. Immunogenetics of systemic sclerosis: defining heritability, functional variants and shared-autoimmunity pathways. *J. Autoimmun.* **64**, 53–65 (2015).
- Zhang, F. et al. Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.* **10**, 451–481 (2009).
- Henrichsen, C. N. et al. Copy number variants, diseases and gene expression. *Hum. Mol. Genet.* **18**, R1–R8 (2009).
- Fanciulli, M. et al. FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat. Genet.* **39**, 721–723 (2007).
- Yang, Y. et al. Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. *Am. J. Hum. Genet.* **80**, 1037–1054 (2007).
- de Cid, R. et al. Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nat. Genet.* **41**, 211–215 (2009).
- McKinney, C. et al. Association of variation in Fcγ receptor 3B gene copy number with rheumatoid arthritis in Caucasian samples. *Ann. Rheum. Dis.* **69**, 1711–1716 (2010).
- Olsson, L. M. et al. Copy number variation in autoimmunity—importance hidden in complexity? *Eur. J. Immunol.* **42**, 1969–1976 (2012).
- Usher, C. L. et al. Complex and multi-allelic copy number variation in human disease. *Brief. Funct. Genomics* **14**, 329–338 (2015).
- Carroll, M. C. The role of complement and complement receptors in induction and regulation of immunity. *Annu. Rev. Immunol.* **16**, 545–568 (1998).
- Walport, M. J. Complement. First of two parts. *N. Engl. J. Med.* **344**, 1058–1066 (2001).
- West, E. E. et al. Complement and the regulation of T cell responses. *Annu. Rev. Immunol.* **36**, 309–338 (2018).
- Chen, M. et al. The complement system in systemic autoimmune disease. *J. Autoimmun.* **34**, J276–J286 (2010).
- Holers, V. M. Complement and its receptors: new insights into human disease. *Annu. Rev. Immunol.* **32**, 433–459 (2014).
- Goicoechea de Jorge, E. et al. Common and rare genetic variants of complement components in human disease. *Mol. Immunol.* **102**, 42–57 (2018).
- Yang, Y. et al. Diversity in intrinsic strengths of the human complement system: serum C4 protein concentrations correlate with C4 gene size and polygenic variations, hemolytic activities, and body mass index. *J. Immunol.* **171**, 2734–2745 (2003).
- Wang, H. et al. Complement C4, infections, and autoimmune diseases. *Front. Immunol.* **12**, 694928 (2021).
- Banlaki, Z. et al. Fine-tuned characterization of RCCX copy number variants and their relationship with extended MHC haplotypes. *Genes Immun.* **13**, 530–535 (2012).
- Wu, Z. et al. Association between complement 4 copy number variation and systemic lupus erythematosus: a meta-analysis. *Clin. Exp. Med.* **20**, 627–634 (2020).
- Okada, Y. et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).
- Langefeld, C. D. et al. Transancestral mapping and genetic load in systemic lupus erythematosus. *Nat. Commun.* **8**, 16021 (2017).
- Kamitaki, N. et al. Complement genes contribute sex-biased vulnerability in diverse disorders. *Nature* **582**, 577–581 (2020).
- Sekar, A. et al. Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).
- Yu, C. Y. et al. Dancing with complement C4 and the RP-C4-CYP21-TNX (RCCX) modules of the major histocompatibility complex. *Prog. Nucleic Acid Res. Mol. Biol.* **75**, 217–292 (2003).
- Klein, M. et al. Contribution of CD8+ T cells to inflammatory cytokine production in systemic sclerosis (SSc). *Autoimmunity* **49**, 532–546 (2016).
- Ingersoll, M. A. Sex differences shape the response to infectious diseases. *PLoS Pathog.* **13**, e1006688 (2017).
- Blasko, B. et al. Low complement C4B gene copy number predicts short-term mortality after acute myocardial infarction. *Int. Immunol.* **20**, 31–37 (2008).
- Walli-Attaei, M. et al. Variations between women and men in risk factors, treatments, cardiovascular disease incidence, and death in 27 high-income, middle-income, and low-income countries (PURE): a prospective cohort study. *Lancet* **396**, 97–109 (2020).
- Liang, Y. et al. A gene network regulated by the transcription factor VGLL3 as a promoter of sex-biased autoimmune diseases. *Nat. Immunol.* **18**, 152–160 (2017).
- Boese, A. C. et al. Sex differences in vascular physiology and pathophysiology: estrogen and androgen signaling in health and disease. *Am. J. Physiol. Heart Circ. Physiol.* **313**, H524–H545 (2017).
- Han, M. K. et al. Female sex and gender in lung/sleep health and disease. increased understanding of basic biological, pathophysiological, and behavioral mechanisms leading to better health for female patients with lung disease. *Am. J. Respir. Crit. Care Med.* **198**, 850–858 (2018).
- Blanchong, C. A. et al. Genetic, structural and functional diversities of human complement components C4A and C4B and their mouse homologues, Slp and C4. *Int. Immunopharmacol.* **1**, 365–392 (2001).
- Kerick, M. et al. Expression quantitative trait locus analysis in systemic sclerosis identifies new candidate genes associated with multiple aspects of disease pathology. *Arthritis Rheumatol.* **73**, 1288–1300 (2021).
- Basu, M. et al. Predicting tissue-specific gene expression from whole blood transcriptome. *Sci. Adv.* **7**, eabd6991 (2021).
- Mu, Z. et al. The impact of cell type and context-dependent regulatory variants on human immune traits. *Genome Biol.* **22**, 122 (2021).
- Gadjeva, M. et al. Macrophage-derived complement component C4 can restore humoral immunity in C4-deficient mice. *J. Immunol.* **169**, 5489–5495 (2002).
- Wouters, D. et al. High-throughput analysis of the C4 polymorphism by a combination of MLPA and isotype-specific ELISA's. *Mol. Immunol.* **46**, 592–600 (2009).
- Juptner, M. et al. Low copy numbers of complement C4 and homozygous deficiency of C4A may predispose to severe disease and earlier disease onset in patients with systemic lupus erythematosus. *Lupus* **27**, 600–609 (2018).
- Esposito, J. et al. The association of low complement with disease activity in systemic sclerosis: a prospective cohort study. *Arthritis Res. Ther.* **18**, 246 (2016).
- Mitchell, T. J. et al. IFN-gamma up-regulates expression of the complement components C3 and C4 by stabilization of mRNA. *J. Immunol.* **156**, 4429–4434 (1996).
- Beretta, L. et al. Genome-wide whole blood transcriptome profiling in a large European cohort of systemic sclerosis patients. *Ann. Rheum. Dis.* **79**, 1218–1226 (2020).
- Agudelo, W. A. et al. Quantum chemical analysis of MHC-peptide interactions for vaccine design. *Mini Rev. Med. Chem.* **10**, 746–758 (2010).
- Kongkaew, S. et al. Interactions of HLA-DR and topoisomerase I epitope modulated genetic risk for systemic sclerosis. *Sci. Rep.* **9**, 745 (2019).
- Acosta-Herrera, M. et al. Comprehensive analysis of the major histocompatibility complex in systemic sclerosis identifies differential HLA associations by clinical and serological subtypes. *Ann. Rheum. Dis.* **80**, 1040–1047 (2021).
- Blanchong, C. A. et al. Deficiencies of human complement component C4A and C4B and heterozygosity in length variants of RP-C4-CYP21-TNX (RCCX) modules in caucasians. The load of RCCX genetic diversity on major histocompatibility complex-associated disease. *J. Exp. Med.* **191**, 2183–2196 (2000).
- Matsumoto, K. I. et al. The roles of tenascins in cardiovascular, inflammatory, and heritable connective tissue diseases. *Front. Immunol.* **11**, 609752 (2020).
- Kasprzycka, M. et al. Tenascins in fibrotic disorders—from bench to bedside. *Cell Adh. Migr.* **9**, 83–89 (2015).
- Valcourt, U. et al. Tenascin-X: beyond the architectural function. *Cell Adh. Migr.* **9**, 154–165 (2015).
- Ricklin, D. et al. The renaissance of complement therapeutics. *Nat. Rev. Nephrol.* **14**, 26–47 (2018).
- Gouin, A. et al. Role of C5 inhibition in idiopathic inflammatory myopathies and scleroderma renal crisis-induced thrombotic microangiopathies. *Kidney Int. Rep.* **6**, 1015–1021 (2021).
- Devresse, A. et al. Complement activation and effect of eculizumab in scleroderma renal crisis. *Medicine* **95**, e4459 (2016).

58. van den Hoogen, F. et al. 2013 classification criteria for systemic sclerosis: an American college of rheumatology/European league against rheumatism collaborative initiative. *Ann. Rheum. Dis.* **72**, 1747–1755 (2013).
59. LeRoy, E. C. et al. Criteria for the classification of early systemic sclerosis. *J. Rheumatol.* **28**, 1573–1576 (2001).
60. Barturen, G. et al. Integrative analysis reveals a molecular stratification of systemic autoimmune diseases. *Arthritis Rheumatol.* **73**, 1073–1085 (2021).
61. Tang, H. et al. Estimation of individual admixture: analytical and study design considerations. *Genet Epidemiol.* **28**, 289–301 (2005).
62. Martin, M. Cutadapt remove adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).
63. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
64. Li, B. et al. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinforma.* **12**, 323 (2011).
65. Huber, W. et al. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18**, S96–S104 (2002).
66. Capaldo, C. et al. The active immunological profile in patients with primary Sjogren's syndrome is restricted to typically encountered autoantibodies. *Clin. Exp. Rheumatol.* **34**, 722 (2016).
67. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
68. Mason, M. J. et al. Low HERV-K(C4) copy number is associated with type 1 diabetes. *Diabetes* **63**, 1789–1795 (2014).
69. Zai, C. C. et al. Association study of the complement component C4 gene in tardive dyskinesia. *Front. Pharm.* **10**, 1339 (2019).
70. Mariaselvam, C. M. et al. The complement C4 genetic diversity in first episode psychosis of the OPTiMISE cohort. *Schizophr Bull Open* **2**, sgab003 (2021).
71. Jia, X. et al. Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS ONE* **8**, e64683 (2013).
72. Brown, W. M. et al. Overview of the MHC fine mapping data. *Diabetes Obes. Metab.* **11**, 2–7 (2009).
73. Han, B. et al. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am. J. Hum. Genet.* **88**, 586–598 (2011).
74. Rucker, J. J. et al. Phenotypic association analyses with copy number variation in recurrent depressive disorder. *Biol. Psychiatry* **79**, 329–336 (2016).
75. Dai, J. Y. et al. Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. *Biometrika* **99**, 929–944 (2012).

## ACKNOWLEDGEMENTS

We would like to thank Guillermo Barturen Briñas and Elena Carnero-Montoro for fruitful discussions and Sofia Vargas and Gema Robledo for their excellent technical assistance. We would like to thank Elena López-Isac for organizing all SSC GWAS datasets and all members of the PRECISEADS consortium, especially Ralf Lesche, Sepideh Babaei, Anne Buttgerit, Suzana Makowska and Martina Runge for preparing the RNA Seq data and Johan Frostegård and Jacques-Olivier Pers for preparing and normalizing the serum C4 data. We greatly appreciate the patients and healthy donors for their essential participation in these studies. This work was supported by grant RTI2018101332-B-I00 funded by MCIN/AEI/10.13039/501100011033 by “ERDF A way of making Europe”, Red de Investigación en Inflamación y Enfermedades

## INTERNATIONAL SSC GROUP

P. Carreira<sup>20</sup>, I. Castellvi<sup>21</sup>, José Luis Callejas<sup>4</sup>, R. Ríos<sup>4</sup>, R. García Portales<sup>22</sup>, A. Fernández-Nebro<sup>23</sup>, F. J. García-Hernández<sup>24</sup>, M. A. Aguirre<sup>25</sup>, B. Fernández-Gutiérrez<sup>26</sup>, L. Rodríguez-Rodríguez<sup>26</sup>, P. García de la Peña<sup>27</sup>, E. Vicente<sup>28</sup>, J. L. Andreu<sup>29</sup>, M. Fernández de Castro<sup>29</sup>, F. J. López-Longo<sup>30</sup>, and Carmen Pilar Simeón-Aznar<sup>3</sup>, V. Fonollosa<sup>3</sup>, A. Guillén<sup>3</sup>, G. Espinosa<sup>31</sup>, C. Tolosa<sup>32</sup>, A. Pros<sup>33</sup>, E. Beltrán<sup>33</sup>, M. Rodríguez Carballeira<sup>34</sup>, F. J. Narváez<sup>35</sup>, M. Rubio Rivas<sup>35</sup>, V. Ortiz-Santamaría<sup>36</sup>, A. B. Madroño<sup>37</sup>, M. A. González-Gay<sup>38</sup>, B. Díaz<sup>39</sup>, L. Trapiella<sup>39</sup>, M. V. Egurbide<sup>40</sup>, P. Fanlo-Mateo<sup>41</sup>, L. Saez-Comet<sup>42</sup>, F. Díaz<sup>43</sup>, J. A. Roman-Ivorra<sup>44</sup>, J. J. Alegre Sancho<sup>45</sup>, M. Freire<sup>46</sup>, F. J. Blanco García<sup>47</sup>, N. Oreiro<sup>47</sup>, T. Witte<sup>48</sup>, A. Kreuter<sup>49</sup>, G. Riemekaster<sup>50</sup>, P. Airo<sup>51</sup>, C. Magro<sup>10</sup>, A. E. Voskuyl<sup>52</sup>, M. C. Vonk<sup>53</sup>, R. Hesselstrand<sup>54</sup>, A. Nordin<sup>55</sup>, C. Lunardi<sup>56</sup>, A. Gabrielli<sup>57</sup>, A. Hoffmann-Vold<sup>58</sup>, J. H. W. Distler<sup>59</sup>, L. Padyukov<sup>55</sup> and B. P. C. Koeleman<sup>60</sup>

<sup>20</sup>Department of Rheumatology, 12 de Octubre University Hospital, Madrid, Spain. <sup>21</sup>Department of Rheumatology, Santa Creu i Sant Pau University Hospital, Barcelona, Spain. <sup>22</sup>Department of Rheumatology, Virgen de la Victoria Hospital, Málaga, Spain. <sup>23</sup>Department of Rheumatology, Carlos Haya Hospital, Málaga, Spain. <sup>24</sup>Department of Internal Medicine, Virgen del Rocío Hospital, Sevilla, Spain. <sup>25</sup>Department of Rheumatology, Reina Sofía/IMBIC Hospital, Córdoba, Spain. <sup>26</sup>Department of Rheumatology, San Carlos Clinic Hospital, Madrid, Spain. <sup>27</sup>Department of Rheumatology, Madrid Norte Sanchinarro Hospital, Madrid, Spain. <sup>28</sup>Department of Rheumatology, La Princesa Hospital, Madrid, Spain. <sup>29</sup>Department of Rheumatology, Puerta de Hierro Hospital-Majadahonda, Madrid, Spain. <sup>30</sup>Department of Rheumatology, Gregorio Marañón University Hospital, Madrid, Spain. <sup>31</sup>Department of Internal Medicine, Clinic Hospital, Barcelona, Spain. <sup>32</sup>Department of Internal Medicine, Parc Tauli Hospital, Sabadell, Spain. <sup>33</sup>Department of Rheumatology, Hospital Del Mar, Barcelona, Spain. <sup>34</sup>Department of Internal Medicine, Hospital Universitari Mútua Terrasa, Barcelona, Spain. <sup>35</sup>Department of Rheumatology, Bellvitge University

Reumáticas (RIER) from Instituto de Salud Carlos III (RD16/0012/0013). This work has received funding from the Innovative Medicines Initiative 1 & 2 Joint Undertaking (JU) under grant agreements No 115565 (PRECISEADS) and No 831434 (3TR). The JU receives support from the European Union's FP7 and Horizon 2020 research and innovation programs and EFPIA. MAH was supported by the Juan de la Cierva Incorporación program, grant IJC2018-035131-I funded by MCIN/AEI/10.13039/501100011033. This work is dedicated to the memory of Annette Kerick (1945–2020).

## AUTHOR CONTRIBUTIONS

M.K., M.A.H., and J.M. contributed to the conception and study design. M.K. and M.A.H. contributed to data collection, QC, imputation, and data analysis. C.P.S.-A., J.L.C., S.A., S.M.P., M.N., N.H., G.M., J.K.V.-B., G.O., A.B., A.H., C.T., Y.A., C.F., M.E.A.-R., T.R.D.J.R., L.B., C.P.D., and M.D.M. contributed to GWAS and RNA-Sequencing data collection. All co-authors made substantial contributions to data acquisition, data interpretation, and revised the work critically for important intellectual content.

## COMPETING INTERESTS

The authors declare no competing interests.

## ADDITIONAL INFORMATION

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41525-022-00327-8>.

**Correspondence** and requests for materials should be addressed to Martin Kerick, Marialbert Acosta-Herrera or Javier Martin.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022

Hospital, Barcelona, Spain. <sup>36</sup>Department of Rheumatology, Granollers Hospital, Granollers, Spain. <sup>37</sup>Department of Internal Medicine, Hospital General San Jorge, Huesca, Spain. <sup>38</sup>Epidemiology, Genetics and Atherosclerosis Research Group on Systemic Inflammatory Diseases, DIVAL, University of Cantabria, Santander, Spain. <sup>39</sup>Department of Internal Medicine, Hospital Central de Asturias, Oviedo, Spain. <sup>40</sup>Department of Internal Medicine, Hospital Universitario Cruces, Barakaldo, Spain. <sup>41</sup>Department of Internal Medicine, Hospital Virgen del Camino, Pamplona, Spain. <sup>42</sup>Department of Internal Medicine, Hospital Universitario Miguel Servet, Zaragoza, Spain. <sup>43</sup>Department of Rheumatology, Hospital Universitario de Canarias, Tenerife, Spain. <sup>44</sup>Department of Rheumatology, Hospital Universitari i Politècnic La Fe, Valencia, Spain. <sup>45</sup>Department of Rheumatology, Hospital Universitari Doctor Peset, Valencia, Spain. <sup>46</sup>Department of Internal Medicine, Thrombosis and Vasculitis Unit, Complejo Hospitalario Universitario de Vigo, Vigo, Spain. <sup>47</sup>Department of Rheumatology, INIBIC-Hospital Universitario A Coruña, La Coruña, Spain. <sup>48</sup>Department of Clinical Immunology, Hannover Medical School, Hannover, Germany. <sup>49</sup>Department of Dermatology, Josefs-Hospital, Ruhr University Bochum, Bochum, Germany. <sup>50</sup>Clinic of Rheumatology, University of Lübeck, Lübeck, Germany. <sup>51</sup>Service of Rheumatology and Clinic Immunology Spedali Civili, Brescia, Italy. <sup>52</sup>Department of Rheumatology, VU University Medical Center, Amsterdam, The Netherlands. <sup>53</sup>Department of Rheumatology, Radboud University Nijmegen Medical Center, Nijmegen, Netherlands. <sup>54</sup>Department of Rheumatology, Lund University, Lund, Sweden. <sup>55</sup>Division of Rheumatology, Department of Medicine, Karolinska University Hospital, Karolinska Institute, Stockholm, Sweden. <sup>56</sup>Department of Medicine, Università degli Studi di Verona, Verona, Italy. <sup>57</sup>Istituto di Clinica Medica Generale, Ematologia ed Immunologia Clinica, Università Politecnica delle Marche, Ancona, Italy. <sup>58</sup>Department of Rheumatology, Oslo University Hospital, Oslo, Norway. <sup>59</sup>Department of Internal Medicine 3, Institute for Clinical Immunology, University of Erlangen-Nuremberg, Erlangen, Germany. <sup>60</sup>Department of Genetics, University Medical Center Utrecht, Utrecht, The Netherlands

### AUSTRALIAN SCLERODERMA INTEREST GROUP (ASIG)

Susanna M. Proudman<sup>6</sup>, Mandana Nikpour<sup>7</sup>, W. Stevens<sup>7</sup>, J. Zochling<sup>61</sup>, J. Sahhar<sup>62</sup>, J. Roddy<sup>63</sup>, P. Nash<sup>64</sup>, K. Tymms<sup>65</sup>, M. Rischmueller<sup>66</sup> and S. Lester<sup>66</sup>

<sup>61</sup>Menzies Research Institute Tasmania, University of Tasmania, Hobart, TAS, Australia. <sup>62</sup>Department Rheumatology, Monash Medical Centre, Melbourne, VIC, Australia. <sup>63</sup>Rheumatology, Royal Perth Hospital, Perth, WA, Australia. <sup>64</sup>Research Unit, Sunshine Coast Rheumatology, Maroochydore, QLD, Australia. <sup>65</sup>Canberra Rheumatology, Canberra, ACT, Australia. <sup>66</sup>Department Rheumatology, The Queen Elizabeth Hospital, Woodville, SA, Australia

### PRECISESADS CLINICAL CONSORTIUM

Lorenzo Beretta<sup>19</sup>, Barbara Vigone<sup>19</sup>, Jacques-Olivier Pers<sup>67</sup>, Alain Saraux<sup>67</sup>, Valérie Devauchelle-Pensec<sup>67</sup>, Divi Cornec<sup>67</sup>, Sandrine Jousse-Joulin<sup>67</sup>, Bernard Lauwerys<sup>68</sup>, Julie Ducreux<sup>68</sup>, Anne-Lise Maudoux<sup>68</sup>, Carlos Vasconcelos<sup>69</sup>, Ana Tavares<sup>69</sup>, Esmeralda Neves<sup>69</sup>, Raquel Faria<sup>69</sup>, Mariana Brandão<sup>69</sup>, Ana Campar<sup>69</sup>, António Marinho<sup>69</sup>, Fátima Farinha<sup>69</sup>, Isabel Almeida<sup>69</sup>, Miguel Angel Gonzalez-Gay Mantecón<sup>38</sup>, Ricardo Blanco Alonso<sup>38</sup>, Alfonso Corrales Martínez<sup>38</sup>, Ricard Cervera<sup>31</sup>, Ignasi Rodríguez-Pintó<sup>31</sup>, Gerard Espinosa<sup>31</sup>, Rik Lories<sup>70</sup>, Ellen De Langhe<sup>70</sup>, and Nicolas Hunzelmann<sup>8</sup>, Doreen Belz<sup>8</sup>, Torsten Witte<sup>48</sup>, Niklas Baerlecken<sup>48</sup>, Georg Stummvoll<sup>71</sup>, Michael Zauner<sup>71</sup>, Michaela Lehner<sup>71</sup>, Eduardo Collantes<sup>72</sup>, Rafaela Ortega-Castro<sup>72</sup>, Ma Angeles Aguirre-Zamorano<sup>72</sup>, Alejandro Escudero-Contreras<sup>72</sup>, Ma Carmen Castro-Villegas<sup>73</sup>, María Concepción Fernández Roldán<sup>73</sup>, Norberto Ortega<sup>74</sup>, Enrique Raya<sup>75</sup>, Inmaculada Jiménez Moleón<sup>75</sup>, Enrique de Ramon<sup>76</sup>, Isabel Díaz Quintero<sup>76</sup>, Pier Luigi Meroni<sup>77</sup>, Maria Gerosa<sup>77</sup>, Tommaso Schioppo<sup>77</sup>, Carolina Artusi<sup>77</sup>, Carlo Chizzolini<sup>78</sup>, Aleksandra Zuber<sup>78</sup>, Donatienne Wynar<sup>78</sup>, Laszlo Kovács<sup>79</sup>, Attila Balog<sup>79</sup>, Magdolna Deák<sup>79</sup>, Márta Bocskai<sup>79</sup>, Sonja Dulic<sup>79</sup>, Gabriella Kádár<sup>79</sup>, Falk Hiepe<sup>80</sup>, Velia Gerl<sup>80</sup>, Silvia Thiel<sup>80</sup>, Manuel Rodriguez Maresca<sup>81</sup>, Antonio López-Berrio<sup>81</sup>, Rocío Aguilar-Quesada<sup>81</sup> and Héctor Navarro-Linares<sup>81</sup>

<sup>67</sup>Centre Hospitalier Universitaire de Brest, Hospital de la Cavale Blanche, Brest, France. <sup>68</sup>Pôle de pathologies rhumatismales systémiques et inflammatoires, Institut de Recherche Expérimentale et Clinique, Université catholique de Louvain, Brussels, Belgium. <sup>69</sup>Centro Hospitalar do Porto, Porto, Portugal. <sup>70</sup>Katholieke Universiteit Leuven, Leuven, Belgium. <sup>71</sup>Medical University Vienna, Vienna, Austria. <sup>72</sup>Servicio Andaluz de Salud, Hospital Universitario Reina Sofía Córdoba, Córdoba, Spain. <sup>73</sup>Servicio Andaluz de Salud, Complejo hospitalario Universitario de Granada (Hospital Universitario San Cecilio), Granada, Spain. <sup>74</sup>Department of Medicine, University of Granada, Granada, Spain. <sup>75</sup>Servicio Andaluz de Salud, Complejo hospitalario Universitario de Granada (Hospital Virgen de las Nieves), Granada, Spain. <sup>76</sup>Servicio Andaluz de Salud, Hospital Regional Universitario de Málaga, Málaga, Spain. <sup>77</sup>Università degli studi di Milano, Milan, Italy. <sup>78</sup>Hospitaux Universitaires de Genève, Genève, Switzerland. <sup>79</sup>University of Szeged, Szeged, Hungary. <sup>80</sup>Charité, Berlin, Germany. <sup>81</sup>Andalusian Public Health System Biobank, Granada, Spain.