Australian Scleroderma Interest Group (ASIG)

J. Zochling[1], J. Sahhar[2], J. Roddy[3], P. Nash[4], K. Tymms[5], M. Rischmueller[6], S. Lester[7]

1. Menzies Research Institute Tasmania, University of Tasmania, Hobart, TAS, Australia.
2. Department Rheumatology, Monash Medical Centre,Melbourne, VIC, Australia.
3. Rheumatology, Royal Perth Hospital, Perth, WA, Australia.
4. Research Unit, Sunshine Coast Rheumatology,Maroochydore, QLD, Australia.
5. Canberra Rheumatology, Canberra, ACT, Australia.
6. Department Rheumatology, The Queen Elizabeth Hospital,Woodville, SA, Australia

PRECISESADS collaborator group

Doreen Belz[1], Francesca Ingegnoli[2] , Yolanda Jimenez Gómez[3] , Chary Lopez Pedrera[3] ,Rik Lories[4] , Eduardo Collantes-Estevez[3] , Gaia Montanelli[5] , Silvia Piantoni[6] , Ignasi Rodriguez Pinto[7], Carlos Vasconcelos[8]

1. Klinik und Poliklinik für Dermatologie und Venerologie, Uniklinik Köln, Köln, Germany
2. Department of Clinical Sciences and Community Health, University of Milan, Milan, Italy
3. IMIBIC, Reina Sofia Hospital, University of Cordoba, Cordoba, Spain
4. Division of Rheumatology, University Hospitals Leuven and Skeletal Biology and Engineering Research Center, KU Leuven, Leuven, Belgium
5. Scleroderma Unit, Referral Center for Systemic Autoimmune Diseases, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico di Milano, Milan, Italy.
6. Immunology & Allergy, University Hospital and School of Medicine, Geneva, Switzerland
7. Division of Rheumatology, University Hospitals Leuven and Skeletal Biology and Engineering Research Center, KU Leuven, Leuven, Belgium
8. Serviço de Imunologia EX-CICAP, Centro Hospitalar e Universitário do Porto, Porto, Portugal

PRECISESADS Flow Cytometry study group

Christophe Jamin[1], Concepción Marañón[2], Lucas Le Lann[1], Quentin Simon[1], Bénédicte Rouvière[1], Nieves Varela[2], Brian Muchmore[2], Aleksandra Dufour[3], Montserrat Alvarez[3], Jonathan Cremer[4], Nuria Barbarroja[5], Velia Gerl[6], Laleh Khodadadi[6], Qingyu Cheng[6], Anne Buttgereit[7], Aurélie De Groof[8], Julie Ducreux[8], Elena Trombetta[9], Tianlu Li[10], Damiana Alvarez-Errico[10], Torsten Witte[11], Katja Kniesch[11], Nancy Azevedo[2], Esmeralda Neves[2], Nancy Azevedo[5,12], Esmeralda Neves[5,12], Sambasiva Rao[13], Pierre-Emmanuel Jouve[14].

1. 1U1238, Université de Brest, Inserm, Labex IGO, CHU de Brest, Brest, France
2. GENYO, Centre for Genomics and Oncological Research Pfizer, University of Granada, Andalusian Regional Government, PTS GRANADA, Granada, Spain
3. Immunology & Allergy, University Hospital and School of Medicine, Geneva, Switzerland
4. Laboratory of Clinical Immunology, Department of Microbiology and Immunology, KU Leuven, Leuven, Belgium
5. IMIBIC, Reina Sofia Hospital, University of Cordoba, Cordoba, Spain
6. Department of Rheumatology and Clinical Immunology, Charité University Hospital, Berlin, Germany
7. Bayer AG, Berlin, Germany
8. Pôle de Pathologies Rhumatismales Inflammatoires et Systémiques, Institut de Recherche Expérimentale et Clinique, Université catholique de Louvain, Brussels, Belgium

9. Laboratorio di Analisi Chimico Cliniche e Microbiologia - Servizio di Citofluorimetria, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico di Milano, Milano, Italy
10. Chromatin and Disease Group, Bellvitge Biomedical Research Institute (IDIBELL), Barcelona, Spain
11. Klinik für Immunologie und Rheumatologie, Medical University Hannover, Hannover, Germany;
12. Serviço de Imunologia EX-CICAP, Centro Hospitalar e Universitário do Porto, Porto, Portugal
13. Sanofi Genzyme, Framingham, MA, USA
14. AltraBio SAS, Lyon, France

SUPPLEMENTARY METHODS

Cohort description

We used the summary statistics from the SSc meta-GWAS by Lopez-Isac et al. [1]. This study comprised 9,095 patients with SSc and 17,584 healthy controls, from 14 cohorts of European ancestry (**Supplementary Figure 1**), but different geographical origin (including Europe, USA and Australia). The composition of the cohort is described in the original data set [1].

In brief, quality control (QC) thresholds for all GWAS datasets were applied as follows: SNPs with call rates < 0.98, minor allele frequencies (MAFs) < 0.01 and deviations from Hardy-Weinberg equilibrium (HWE; p < 0.001 in both case and control subjects) were discarded from further analysis. To perform the PC analysis and identification of outliers, PLINK [2], GCTA [3] and R-base software under GNU Public license v.2 were used. Samples showing > 4 standard deviations from the cluster centroids of each cohort were considered outliers and removed from further analyses.

For the imputation, the IMPUTE2 [4] software was used, adapting the data to the previous format with GTOOL following the QC described above. The 1000 Genome Project Phase 3 (1KGPh3) [5] was used as a reference panel. 4,723,365 SNPs were left for further analysis after QC and imputation. PLINK [2] software was used for the association analysis of the whole genome and to perform a meta-analysis of the fixed effect variance, as well as to calculate the heterogeneity of the odds ratios (ORs) between the cohorts. The level of significance of the associations was established at a value of p ≤ 5 × 10-8.

The deeply phenotyped cohort recruited in the PRECISESADS project [6] was used as a score development cohort. This cohort was composed of 571 healthy controls, 400 patients with SSc, 428 patients with systemic lupus erythematosus (SLE), 380 patients with rheumatoid arthritis (RA), and 395 patients with Sjögren syndrome (SJS). The Illumina HumanCore-24 v1.0 and the Infinium CoreExome-24 v1.2 genome-wide SNP genotyping platform (Illumina Inc., San Diego, CA, USA) were used. After QC the genotyped dataset contained 223,463 SNP loci of 2,073 individuals. This set was imputed using Minimac3 against the HRC v1.1 Genomes reference panel using the Michigan Imputation Server platform. Genotypes were filtered after imputation to have HWE p > 0.001, MAF > 1 % and imputation info score > 0.7 and resulted in 6,664,685 imputed genotypes [7]. The diagnostic criteria for each disease were based on gold-standard clinical guidelines [8–11]. Patients were included in the dcSSc or lcSSc clinical subgroups; those with definite SSc without fibrotic skin disease and puffy fingers were classified as lcSSc. The anti-centromere (ACA) and anti-topoisomerase (ATA) autoantibodies were determined by standard means. High-resolution computed tomography (HRCT) was used to detect the presence of interstitial abnormalities diagnosed as pulmonary fibrosis.

A comprehensive demographic and immunologic characterization was available for a subset of the testing cohort that comprised 333 patients with SSc and 521 controls (**Supplementary Table 1**). The immune cell fractions were estimated by flow cytometry as described in Jamin, C. et al. [12]

LD Clumping

Linkage disequilibrium (LD) based clumping was performed on the genotypes for both patients with SSc and controls in the GWAS [1] using a 2Mb window with the index SNP at the center (--clump-kb 1M), $R^2 > 0.1$ (--clump-r2 0.1) and no p-value threshold (--clump-p 1). Following this strategy, out of the initial set of 5,360,026 polymorphisms, a total of 131,132 SNPs remained. Out of the latter, 103,842 SNPs (79%) were genotyped or imputed in the PRECISESADS cohort. Furthermore, 52 out of 58 (90%) GWAS-significant SNPs in the clumped GWAS set were available in the PRECISESADS dataset.

The HLA region is a highly polymorphic region that presents complex LD patterns. Therefore, in order to avoid an overrepresentation of the HLA in the GRS, all variants in the HLA region (chr6:20,000,000-40,000,000) were discarded from the analysis and only the most associated SNP in the HLA for the SSc vs. controls (rs6457617), the dcSSc vs. lcSSc (rs9275332) or ATA+ vs. ACA+ (rs9275372) comparisons were included in the corresponding GRS. Therefore, the final set of analyzed variants comprised a total of 129,918 SNPs.

GRS additive model description and details

For GRS calculation using PRSice-2 [13], we specified an additive genetic model. Therefore, G was defined as the number of effective alleles observed (G = 0, G = 1 or G = 2, respectively) for the ith genetic variant, and S corresponded to the summary statistic for the effective allele. We also specified the "--score avg" option, and consequently the GRS was calculated as described in PRSice-2:

$$GRSj = \Sigma \frac{SixGij}{Mj}$$

Considering Mj as the number of alleles included in the GRS for the jth individual.

Model-fitting analyses

Model fit: R2 of the full model (SSc case or control ~ GRS + Sex) - R2 of the null model (SSc case orcontrol ~ Sex). The best fitting model was selected based on the highest variance explained (R2). We calculated the Lee R2 [14], which is applied on the liability scale here and it estimates the proportion of variance explained by the GRS of a hypothetical normally distributed latent variable that underlies and causes case/control status.

Two-sided Student's t-tests were performed using the "t.test" function implemented in the "stats v3.6.2" R package [15].

Basic receiver operating characteristic (ROC) / area under the curve (AUC) analyses were performed by using the pROC package [16]. ROC curve properties were defined as follows:

Sensitivity = True Positives / (True Positives + False Negatives)

Specificity = True Negatives / (True Negatives + False Positive)

Accuracy = (True Negatives + True Positives) / (True Negatives + True Positives + False Negatives + False Positives)

Generalized linear models were fitted using the "glm" function and likelihood-ratio test (LRT) p-values were obtained using the “anova” function (test = “LRT”, option was specified) as implemented in the “stats v3.6.2” R package [15].

AUC correlation with country latitude, longitude and distance to European reference populations

Country latitudes and longitudes were obtained from: https://developers.google.com/public-data/docs/canonical/countries_csv. Principal components were calculated using PLINK [2] (--pca option), then distances from each individual to the CEU+GBR population centroid in the top 2 PCs projections were calculated. Afterwards, these parameters were correlated to country-specific AUC values by using linear models (lm function as implemented in the “stats v3.6.2” R package [15].

We observed that including the cohort of origin as a covariate did not affect the p-value threshold selection, but contributed to the proportion of variance explained by the model (full model $R^2 = 0.24$).

AUC p-value calculation

95% confidence intervals were calculated using the ci=TRUE option for the roc() function implemented in pROC. Then, the variance was obtained using pROC::var*, which allowed us to calculate a standard error manually. Finally, a p-value based on the Z-distribution was calculated (two-sided test).

Bibliography

1 López-Isac E, European Scleroderma Group†, Acosta-Herrera M, *et al.* GWAS for systemic sclerosis identifies multiple risk loci and highlights fibrotic and vasculopathy pathways. Nature Communications. 2019;**10**. doi:10.1038/s41467-019-12760-y

2 Purcell S, Neale B, Todd-Brown K, *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics. 2007;**81**:559–75. doi:10.1086/519795

3 Yang J, Lee SH, Goddard ME, *et al.* GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 2011;**88**:76–82.

4 Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 2009;**5**:e1000529.

5 Voight BF, Trynka G, B. Howie, C. Fuchsberger, M. Stephens, J. Marchini, GR. Abecasis, *et al.* A global reference for human genetic variation. *Nature* 2015;**526**:68–74.

6 Molecular Reclassification to Find Clinically Useful Biomarkers for Systemic Autoimmune Diseases: Inception Cohort (PRECISESADSI). clinicaltrials.gov. 2016.https://clinicaltrials.gov/ct2/show/NCT02890134

7 Barturen G, Babaei S, Català-Moll F, *et al.* Integrative Analysis Reveals a Molecular Stratification of Systemic Autoimmune Diseases. *medRxiv* 2020;:2020.02.21.20021618.

8 van den Hoogen F, Khanna D, Fransen J, *et al.* 2013 classification criteria for systemic sclerosis: an American college of rheumatology/European league against rheumatism collaborative initiative. *Ann Rheum Dis* 2013;**72**:1747–55.

9 Hochberg MC. Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus. *Arthritis Rheum* 1997;**40**:1725.

10 Aletaha D, Neogi T, Silman AJ, *et al.* 2010 rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. *Ann Rheum Dis* 2010;**69**:1580–8.

11 Vitali C. Classification criteria for Sjogren's syndrome: a revised version of the European criteria proposed by the American-European Consensus Group. Annals of the Rheumatic Diseases. 2002;**61**:554–8. doi:10.1136/ard.61.6.554

12 Jamin C, Le Lann L, Alvarez-Errico D, *et al.* Multi-center harmonization of flow cytometers in the context of the European 'PRECISESADS' project. *Autoimmun Rev* 2016;**15**:1038–45.

13 Choi SW, O'Reilly P. PRSice 2: POLYGENIC RISK SCORE SOFTWARE (UPDATED) AND ITS APPLICATION TO CROSS-TRAIT ANALYSES. European Neuropsychopharmacology. 2019;**29**:S832. doi:10.1016/j.euroneuro.2017.08.092

14 Lee SH, Goddard ME, Wray NR, *et al.* A better coefficient of determination for genetic profile analysis. *Genet Epidemiol* 2012;**36**:214–24.

15 Website. R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/. (accessed 20 May 2020).

16 Robin X, Turck N, Hainard A, *et al.* pROC: an open-source package for R and S to analyze and compare ROC curves. BMC Bioinformatics. 2011;**12**. doi:10.1186/1471-2105-12-77

SUPPLEMENTARY FIGURES

Supplementary Figure 1. Principal component analysis of the genetic data of the score development cohort and the EUR populations in the 1k Genomes project.

Supplementary Figure 2. A) ROC curves for the predictive value of the 33 SNP SSc GRS to distinguish between SSc patients and healthy controls in the score development cohort depending on the cohort of origin. Correlations between the AUC for the SSc GRS and **B)** Geographical longitude of the country **C)** Geographical latitude of the country **D)** Distance of the cohort to the CEU and GBR populations in the 1k Genomes Project.

Supplementary Figure 3. A) ROC curves for the predictive value of the 33 SNP SSc GRS to distinguish between SSc patients in the different clinical subtypes (pink), serological subtypes (purple) and with/without lung fibrosis (light blue) **B)** ROC curves for the predictive value of the clinical subtype-specific GRS (light blue) and the serological subtype-specific GRS (dark blue) to distinguish between SSc patients with/without lung fibrosis.