Supplementary Materials and Methods and Legends of the Supplementary Figures

Supplementary Materials and Methods

Autoantibodies

All autoantibodies were determined at a single center located in Brest, France, between March 2016 and June 2019. The testing for anti-ENAs (extractable nuclear antigens) followed a two-step approach, where all patients underwent an initial screening using the IDS-ISYS chemiluminescent immunoanalyser (Immunodiagnostic, Boldon, UK). If this screening was positive, subsequent specific testing for anti-Scl70 (SCL) autoantibodies was carried out using the same system, along with the testing for the other anti-ENA target antibodies (anti-Sm, -U1-RNP, -Ro52, -Ro60 and -SSB) but not for anti-RNA pol III antibodies. In parallel, anti-centromere antibodies (ACA) were measured similarly for all patients. The established positivity threshold for both ACA and SCL70 autoantibodies was set at 10 arbitrary units (UA/ml) according to the manufacturer's instructions and internal validation. Additionally, antinuclear antibodies (ANAs) were detected by indirect immunofluorescence on HEp-2 cells (ATCC strain: CCL23) using an in-house method. Each sample underwent testing at five successive dilutions (1:80, 1:160, 1:320, 1:640, 1:1280), with the positivity threshold set at 1:160, based on previous recommendations [1] and internal validation.

Immunophenotyping

Immunophenotyping was performed after blood collection in Duraclone tubes (Beckman Coulter) specifically designed and optimized for the PRECISESADS study.

Harmonization of instruments used for flow cytometry analysis was described elsewhere, providing evidence of high-reproducible data in different centers [2,3]. Four distinct panels were used in the study: a general panel including anti-CD16, anti-CD123, anti-HLADR, anti-CD11c, anti-CD15, anti-Lineage (Lin: CD3, CD19, CD56, CD14), a lymphocyte panel including anti-CD56, anti-CD14, anti-CD19, anti-CD3, anti-CD4 and anti-CD8, a B panel (anti-CD19, anti-CD24, anti-CD38, anti-CD27, anti-IgD, anti-TACI, anti-CD5 and anti-CD11b), and a T panel (anti-CD57, anti-CD45RA, anti-CD62L, anti-CD27, anti-CD38, anti-CD3, anti-CD4 and anti-CD8). DC was defined as $HLADR^{+}Lin^{-}CD11c^{+}CD123^{-}$. The geom-bloxplot function from ggplot2 package (v3.5.1) was used for plotting distribution of the data showing the median and error bars representing the ± interquartile range (IQR).

Cytokines

Serum samples were used to measure cytokine levels, including, FAS Ligand, GDF15, CXCL10/IP-10, CCL8/MCP-2, CCL13/MCP-4, CCL4/MIP-1β, CCL17/TARC and IL1-RA, employing the Luminex system. A customized panel of 12 analytes was constructed using a human premixed multi-analyte Luminex assay from R&D Systems. Samples were thawed and analyzed in batches on the same day. Soluble CRP, TNFα, IL-6, BAFF, and TGFβ were measured using commercial ELISA assays as previously described [4]. The geom-bloxplot function from ggplot2 package (v3.5.1) was used for plotting the distribution of the data showing the median and error bars representing the ± interquartile range (IQR)

RNA-Sequencing

RNA sequencing (RNA Seq) data were collected from 115 ACA patients and 91 SCL70 patients with the whole blood RNA-seq data produced by Bayer. The bcl2fastq2 Conversion Software v2.20 was utilized for demultiplexing and converting BCL files. Quality control was ensured using FastQC v0.11.18, with adapter sequences removed by Cutadapt v1.18. Transcriptome alignment was performed with STAR v2.5.2b using the hg19 genome reference (GENCODE v19 annotation), and RSEM v1.2.31 was employed to calculate read counts. Read counts were normalized using the trimmed mean of M-values (TMM) method from the EdgeR package [5] and genes with zero values in all samples were removed. Differential gene expression analysis between groups of patients was conducted using an empirical Bayes method, with a significance threshold set at a fold change $>|1.5|$ and an adjusted p-value <0.05. The fold change threshold of 1.5 was selected to highlight differences between two patient groups exhibiting the same disease. The pathway analysis was performed using the ShinyGo platform v0.77 (absolute fold change ≥ 1.5, FDR cutoff 0.05) [6].

Gene Modules Analysis

Gene expression across 16 distinct conditions including autoimmune and infectious diseases, primary immune deficiencies, cancers, and pregnancy, encompassing 985 unique transcriptome profiles, was analyzed using the BloodGen3Module R package by Rinchai et al [7]. This analysis included a total of 14,168 transcripts. A set of 382 gene co-expression modules was refined to 38 variables (A1 to A38) annotated with pathway, ontology, and literature terms. The fingerprint grid plots generated display the percentage of genes exhibiting a change in transcript abundance, with significant

changes marked by a fold change threshold of >1.5 and an adjusted p-value <0.05. The module annotations, represented as red (increase) or blue (decrease) spots, were derived from the BloodGen3Module R package, with the grid plot shaded grey for non-significant modules and white when no consensus annotation could be ascribed.

Metabolomic Profiling

Metabolomic analysis of serum samples was conducted using an Agilent 1260 HPLC system coupled with a Jet Stream dual ESI interface and an Agilent 6540 Ultra High Definition (UHD) Accurate Mass Q-TOF spectrometer. The MassHunter Profinder Software (B.06.00) was used for the extraction, alignment, and integration of chemical features. Metabolite identification was achieved by comparing the accurate mass, isotopic distribution, and MS/MS fragmentation patterns with existing databases. Statistical analysis was carried out using the omu R package, where the assign hierarchy function was used on the normalized data, followed by the omu_summary function to calculate fold changes and adjusted p-values, with the student's test being the selected statistical test. Pathway analysis was executed using MetaboAnalyst 5.0 software [8]. The geom-bloxplot function from ggplot2 package (v3.5.1) was used for plotting the distribution of the data.

Logistic Regression

For each clinical outcome among skin fibrosis, pulmonary fibrosis, and arthritis, we performed two logistic regressions using the scikit-learn module (v1.5.0). One was

performed with auto antibody status in combination with the identified features (green curve) and one using identified features only (blue curve). AUCs were computed and compared to the predictive value of autoantibody only (orange curve) and features' contributions were extracted from the models in the form of the absolute value of the features' coefficients.

REFERENCES

1 Sack U, Conrad K, Csernok E, *et al.* Autoantibody detection using indirect immunofluorescence on HEp-2 cells. *Ann N Y Acad Sci*. 2009;1173:166–73. doi: 10.1111/j.1749-6632.2009.04735.x

2 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. doi: 10.1016/j.autrev.2016.07.034

3 Le Lann L, Jouve P-E, Alarcón-Riquelme M, *et al.* Standardization procedure for flow cytometry data harmonization in prospective multicenter studies. *Sci Rep*. 2020;10:11567. doi: 10.1038/s41598-020-68468-3

4 Soret P, Le Dantec C, Desvaux E, *et al.* A new molecular classification to drive precision treatment strategies in primary Sjögren's syndrome. *Nat Commun*. 2021;12:3523. doi: 10.1038/s41467-021-23472-7

5 Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26:139–40. doi: 10.1093/bioinformatics/btp616

6 Ge SX, Jung D, Yao R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics*. 2020;36:2628–9. doi: 10.1093/bioinformatics/btz931

7 Rinchai D, Roelands J, Toufiq M, *et al.* BloodGen3Module: blood transcriptional module repertoire analysis and visualization using R. *Bioinformatics*. 2021;37:2382–9. doi: 10.1093/bioinformatics/btab121

8 Tiffany CR, Bäumler AJ. omu, a Metabolomics Count Data Analysis Tool for Intuitive Figures and Convenient Metadata Collection. *Microbiol Resour Announc*. 2019;8:e00129-19. doi: 10.1128/MRA.00129-19

Legends of the Supplementary Figures

Supplementary Figure 1: The transcriptomic analysis comparing SCL70 versus ACA.

(**A**) Expression levels of each DEG in every patient are visualized in a heatmap for each Ab group after centering and scaling the data. Each column corresponds to a patient, and each row represents a DEG. The annotation at the top of the heatmap distinguishes the groups: orange for SCL70 patients, blue for ACA patients, red for treated patients and green for untreated. (**B**) Gene Ontology analysis was performed on the upregulated DEGs identified by comparing the transcriptomic data of SCL70 patients to ACA patients. The fold enrichment for each pathway is presented as -log10(FDR), along with the number of genes associated. **(C)** The BloodGen3Module R package, developed by Rinchai et al. defined 382 modules covering 14,168 transcripts based on co-clustering observations across 16 different states, including autoimmune and infectious diseases, primary immune deficiencies, cancer, and pregnancy, representing 985 unique transcriptome profiles. Within these 382 modules (gene sets), a reduced set of 38 aggregate modules (A1 to A38) was created, consisting of module sets and functionally annotated pathways, ontologies, and literature term enrichments. The annotated fingerprint grid plots show the percentage of genes in each module with increased expression (red spot), decreased expression (blue spot), or no change (white spot), applying thresholds of an absolute fold change cutoff of >1.5 and a false discovery rate-adjusted p-value of <0.05. The BloodGen3Module R package also provides a module grid annotation (available at https://github.com/Drinchai/BloodGen3Module) where color keys are associated with the functional associations of the modules on the grid. When no consensus annotation was attributed, the position on the annotation grid was

colored in white, and the grey corresponded to a position with no module attributed. (**D**) Expression levels of each DEG are visualized in a heatmap for each Ab group only in untreated patients after centering and scaling the data. Each column corresponds to a patient, and each row represents a DEG. The annotation at the top of the heatmap distinguishes the groups: light red for untreated anti-SLC70-positive patients and light blue for untreated anti-ACA-positive patients. (**E**) Gene Ontology analysis was performed on the upregulated DEGs identified by comparing the transcriptomic data of untreated anti-ACA-positive patients to untreated anti- SLC70-positive patients. The fold enrichment for each pathway is presented as -log10(FDR), along with the number of genes associated.

Supplementary Figure 2: Immune cells count and frequencies in in serological groups.

(**A**) Absolute counts (number of cells/mm3) of PMNs, monocytes, DCs, lymphocytes, B cells, $CD4^+$ T cell and CD8 T cells were assessed by flow cytometry in patients and healthy donors (HD) (in grey). (**B**) Frequencies (in $CD3^+$ cells) of $CD4^+$ and $CD8^+$ T cells were analyzed in each group with frequencies of naïve, central memory, effector memory, and TEMRA assessed in $CD4^+$ and $CD8^+$ population in the two groups. SLC70 patients are shown in orange, and ACA patients are in blue (SCL70: 10, ACA: 16). (**C**) Sensitivity comparison of blood cell counts in untreated patients. Light red for untreated anti-SLC70-positive patients light blue for untreated anti-ACA-positive patients and grey for HD. Statistical significance was determined using a t-test and expressed as p-values.

Supplementary Figure 3: Cytokine data in serological groups

(A) Sensitivity comparison of cytokine measurement in untreated patients. IL-6, CCL17/TARC, GDF15, FAS Ligand, and TGFβ were quantified in units in ACA and SCL70 patients as well in healthy donors (HD) (Supplementary Table 5). Colors are light red for untreated anti-SLC70-positive patients light blue for untreated anti-ACA-positive patients and grey for HD. (**B**) Distribution of the cytokine values for IL-6, CCL17 and GDF15 in both groups. Treated patients are visualized in red. Statistical significance was determined using a t-test and expressed as p-values

Supplementary Figure 4: Metabolic profile and metabolite enrichment pathways

(**A**) Concentrations of Indoleacrylic acid, Phosphatidylcholine (14:0), Roemerine, N-1-Deoxy-1-fructosyl leucine and Ethyl aconinate were compared using t-test. Error bars represent the standard deviation (**B-C**) An overview of metabolite set enrichment for the top 25 differentially expressed metabolites was performed using MetaboAnalyst 3.0 software. SCL70 patients are in orange, ACA patients are in blue, and healthy donors (HD) are in grey.

Supplementary Figure 5: Metabolic data in serological group patients

Distribution of all the metabolite values in both groups, with treated patients visualized in red.