**Supplementary Methods**

**Systematic search and data collection**

We used the keywords “systemic sclerosis”, “skin”, “transcriptomics or microarray”, “dataset” in Google Scholar and PubMed to find relevant publications to the topic of synovial gene signatures of patients with SSc (**Figure 1**). We retrieved all publications that were accompanied by high-throughput datasets (13 studies in total). From the resulting set, we removed entries that had been duplicated and selected datasets measuring over 15,000 genes to secure the largest size of genes and samples. Since there was a trade-off between the number of studies to include and the number of genes that are within the intersection from all datasets, we optimized the product of the two by selecting the point where these two trends cross (**Supplementary Figure S1**). The final number of SSc patients was 175 (diffuse type 115, limited type 60) and 61 normal healthy controls (NC) was included.

**The SSc compendium**

The total sample count was 344 in SSc (263 in diffuse type [dSSc] and 81 in limited type [lSSc]) and 79 in NC (**Supplementary Figure S2**). Ultimately, the final SSc compendium was constructed out of 8 studies with a total of 423 samples in 236 persons, covering 23,684 genes total (common core of 17,424 genes). In the SSc compendium, 324 samples were taken from forearm (257 from the involved area of SSc) and 99 from back (87 from the uninvolved area of SSc). To hold the consistency, we filtered only the involved forearm areas and the uninvolved back area in SSc patients for comparison. In 77 patients, 125 samples were harvested after a certain period (range, 6–24 months) of specific drug therapy (rituximab, mycophenolate, nilotinib, and abatacept). Some information on clinical characteristics including presence of interstitial lung disease, pulmonary hypertension, seropositivity of autoantibodies were not fully annotated for each SSc patient (**Supplementary Table S1**).

**Data normalization and removal of batch effects**

For one-channel arrays, the Robust Multi-array Average (RMA) method was applied on the image data for a set of replicates for background correction, normalization, probe-set summarization. For dual-channel arrays, the image data were imported and background correction was performed using normexp as it was shown to outperform other methods.1 Red and green channels were separated and quantile-normalized for each set of replicates. The vectors for the matrices were normalized using the quantile normalization method.2 Residual technical batch effects arising due to heterogeneous data integration were corrected using the ComBat function.2 3 Quality assurance and distribution bias was evaluated by Principal Component Analysis (**Supplementary Figure S3**). After preprocessing, the gene expression profiles have a significant reduction of systematic, dataset-specific bias in comparison with the same dataset before normalization and batch correction (**Supplementary Figure S3**).

**Filtering of differentially expressed genes**

In order to identify the differentially expressed genes (DEGs), we employed three widely-used methods: (a) an empirical Bayesian method using the Benjamini and Hochberg procedure with a signiﬁcance threshold at an adjusted *P*-value < 0.014; (b) the Significance Analysis of Microarray (SAM) method, with a signiﬁcance threshold of false discovery rate FDR < 0.01; (c) the Rank Products (RP) method with a signiﬁcance threshold set at percentage of false prediction pfp < 0.01. The resulting list of DEGs is the intersection of the three individual DEGs sets for each method to minimize the false discovery rate statistic.

**Functional enrichment analysis**

We performed functional enrichment analysis focusing on the list of up-regulated DEGs using the Enrichr software.5 Gene ontology (GO) terms were regarded significant if the adjusted *P*-value is lower than 0.01, and the z-score was less than -3.0. GO terms irrelevant to skin tissue were filtered out.

**Gene set enrichment analysis**

Gene set enrichment analysis (GSEA) analysis was carried out using the GSEA software from the Broad Institute to assess the overrepresentation of SSc-related gene sets.6 The enrichment results were visualized with the Enrichment Map format, where nodes represent gene-sets and weighted links between the nodes represent an overlap score depending on the number of genes two gene-sets share (Jaccard similarity coefficient).7 To intuitively identify redundancies between gene sets, the nodes were connected if their contents overlap by more than 25%. Clusters map to one or more functionally enriched groups, which were manually circled and assigned a label. To test for gene enrichment in individual samples from independent testing dataset (GSE106358), we used a single sample version of gene-set enrichment analysis, which defines an enrichment score as the degree of absolute enrichment of a gene set in each sample within a given data set.8 The gene expression values for a given sample were rank-normalized, and an enrichment score was produced using the Empirical Cumulative Distribution Functions (ECDF) of the genes in the signature and the remaining genes. This procedure is similar to the GSEA technique, but the list is ranked by absolute expression in one sample.

# Construction of protein-protein interaction network

To assess the interconnectivity of DEGs in the SSc skin samples, we constructed a protein-protein network based on the human interactome database.9 In the network, nodes and edges represent genes and functional or physical relationships between them, respectively. Graph theory concepts such as degree, closeness, and betweenness were employed to assess the topology of this network. Hub molecules were defined as the shared genes in top 10% with the highest rank in each arm of the three centrality parameters.10

# Scoring of pathway activation

To quantify certain biological pathway activity, we calculated the gene expression z-scores.11 12 Briefly, a *Z*-score is defined as the difference between the error-weighted mean of the expression values of the genes in each pathway and the error-weighted mean of all genes in a sample after normalization. SSc-associated pathways or processes were curated from the publicly available literatures13-17 and their gene sets were imported from Kyoto Encyclopedia of Genes and Genomes (KEGG)18 and Reactome database.19 *Z*-scores were computed using each pathway in the signature collection for each of the samples, resulting in a matrix of pathway activation scores.

# Non-negative matrix factorization and determination of the optimal number of clusters

To classify the SSc patients into subgroups based on their molecular or pathway signatures, we used the non-negative matrix factorization (NMF) method. NMF clustering is a powerful unsupervised approach to identify the disease subtype or patient subgroup and discover biologically meaningful molecular pattern.12 20 We applied the consensus NMF clustering method and initialized 1000 times for each rank *k* (range from 2 to 6), where k was a presumed number of subtypes in the dataset. For each *k*, 1000 matrix factorizations were used to classify each sample 1000 times. The consensus matrix was used to assess how consistently sample-pairs cluster together. We then computed the cophenetic coefficients and silhouette scores for each *k*, to quantitatively assess global clustering robustness across the consensus matrix. The maximum peak of the cophenetic coefficient and silhouette score plots determined the optimal number of clusters.20 To confirm unsupervised clustering results, we used *t*-distributed stochastic neighborhood embedding (*t*-SNE),21 a powerful dimensionality reduction method. The *t*-SNE method captures the variance in the data by attempting to preserve the distances between data points from high to low dimensions without any prior assumptions about the data distribution.

**Classiﬁcation using a 20-pathway classiﬁer**

We constructed a classiﬁer, where a set of predictors consists of 20 pathways, using a naïve Bayes machine learning algorithm. For training the classiﬁer, we used the pathway activation scores and

subgroup labels of the result of the NMF clustering process. We then computed the misclassiﬁcation rate using stratiﬁed 10-fold cross validation. To assess performance, we employed the one-versus-all comparison method.22 Briefly, we adopted a 4-class classiﬁcation as a 2-class classiﬁcation (e.g., cluster 1 versus others) and computed the average area under the receiver operating characteristic (ROC) curves from all 4 of 2-class classiﬁcations. Finally, we applied the 20-pathway classiﬁer to assign subgroups to the samples.

**Inference of cell types in gene expression clusters**

In order to deconvolute the cellular composition of the 4 clusters in our data, we used an algorithm called xCell,23 a powerful machine learning framework trained on the profiles of 64 immune and stroma cell datasets, for generating cell type enrichment scores and adjusting them to cell type proportions.

# Statistical analysis

For continuous distributed data, between-group comparisons were performed using the one-way ANOVA, unpaired *t*-test or Mann-Whitney *U* test. Categorical or dichotomous variables were compared using the chi-squared test or Fisher’s exact test. Correlation analysis between two variables was carried out using Pearson’s or Spearman’s method. To investigate the difference of pathway activation score across the subgroups, we fitted the one-way ANOVA model using logistic regression. All analyses were conducted in *R* (version 3.5.1, The R Project for Statistical Computing, www.r-project.org).

**References**

1 Ritchie ME, Silver J, Oshlack A*, et al.* A comparison of background correction methods for two-colour microarrays. *Bioinformatics* 2007;23:2700-7.

2 Muller C, Schillert A, Rothemeier C*, et al.* Removing Batch Effects from Longitudinal Gene Expression - Quantile Normalization Plus ComBat as Best Approach for Microarray Transcriptome Data. *PLoS One* 2016;11:e0156594.

3 Chen C, Grennan K, Badner J*, et al.* Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS One* 2011;6:e17238.

4 Datta S, Datta S. Empirical Bayes screening of many p-values with applications to microarray studies. *Bioinformatics* 2005;21:1987-94.

5 Kuleshov MV, Jones MR, Rouillard AD*, et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 2016;44:W90-7.

6 Subramanian A, Tamayo P, Mootha VK*, et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005;102:15545-50.

7 Merico D, Isserlin R, Stueker O*, et al.* Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One* 2010;5:e13984.

8 Barbie DA, Tamayo P, Boehm JS*, et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* 2009;462:108-12.

9 Menche J, Sharma A, Kitsak M*, et al.* Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science* 2015;347:1257601.

10 Koschutzki D, Schreiber F. Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene Regul Syst Bio* 2008;2:193-201.

11 Levine DM, Haynor DR, Castle JC*, et al.* Pathway and gene-set activation measurement from mRNA expression data: the tissue distribution of human pathways. *Genome Biol* 2006;7:R93.

12 You S, Knudsen BS, Erho N*, et al.* Integrated Classification of Prostate Cancer Reveals a Novel Luminal Subtype with Poor Outcome. *Cancer Res* 2016;76:4948-58.

13 Besliu A, Banica L, Predeteanu D*, et al.* Peripheral blood lymphocytes analysis detects CD100/SEMA4D alteration in systemic sclerosis patients. *Autoimmunity* 2011;44:427-36.

14 Bhattacharyya S, Wei J, Varga J. Understanding fibrosis in systemic sclerosis: shifting paradigms, emerging opportunities. *Nat Rev Rheumatol* 2011;8:42-54.

15 Ho YY, Lagares D, Tager AM*, et al.* Fibrosis--a lethal component of systemic sclerosis. *Nat Rev Rheumatol* 2014;10:390-402.

16 Yokoyama Y, Sekiguchi A, Fujiwara C*, et al.* Inhibitory regulation of skin fibrosis in systemic sclerosis by apelin/APJ signaling. *Arthritis Rheumatol* 2018.

17 Yoshizaki A, Yanaba K, Yoshizaki A*, et al.* Treatment with rapamycin prevents fibrosis in tight-skin and bleomycin-induced mouse models of systemic sclerosis. *Arthritis Rheum* 2010;62:2476-87.

18 Kanehisa M, Furumichi M, Tanabe M*, et al.* KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2017;45:D353-d61.

19 Fabregat A, Sidiropoulos K, Garapati P*, et al.* The Reactome pathway Knowledgebase. *Nucleic Acids Res* 2016;44:D481-7.

20 Brunet JP, Tamayo P, Golub TR*, et al.* Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A* 2004;101:4164-9.

21 Maaten LVD, Hinton GE. Visualizing Data using t-SNE. *J Machine Learning Res* 2008;9:2579-605.

22 Patel AC, Markey MK. Comparison of three-class classification performance metrics: a case study in breast cancer CAD. *Medical Imaging*: SPIE, 2005:9.

23 Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol* 2017;18:220.