Distinct molecular subtypes of systemic sclerosis and gene signature with diagnostic capability

Qi Wang[1,2†], Chen-Long Li[1,2†], Li Wu[1,3], Jing-Yi Hu[4], Qi Yu[2,4], Sheng-Xiao Zhang[5]* and Pei-Feng He[2,4]*

[1]School of Basic Medical Sciences, Shanxi Medical University, Taiyuan, China, [2]Shanxi Key Laboratory of Big Data for Clinical Decision Research, Taiyuan, China, [3]Department of Anesthesiology, Shanxi Provincial People's Hospital (Fifth Hospital) of Shanxi Medical University, Taiyuan, China, [4]School of Management, Shanxi Medical University, Taiyuan, China, [5]Department of Rheumatology, Second Hospital of Shanxi Medical University, Taiyuan, China

Background: As Systemic Sclerosis (SSc) is a connective tissue ailment that impacts various bodily systems. The study aims to clarify the molecular subtypes of SSc, with the ultimate objective of establishing a diagnostic model that can inform clinical treatment decisions.

Methods: Five microarray datasets of SSc were retrieved from the GEO database. To eliminate batch effects, the combat algorithm was applied. Immune cell infiltration was evaluated using the xCell algorithm. The ConsensusClusterPlus algorithm was utilized to identify SSc subtypes. Limma was used to determine differential expression genes (DEGs). GSEA was used to determine pathway enrichment. A support vector machine (SVM), Random Forest(RF), Boruta and LASSO algorithm have been used to select the feature gene. Diagnostic models were developed using SVM, RF, and Logistic Regression (LR). A ROC curve was used to evaluate the performance of the model. The compound-gene relationship was obtained from the Comparative Toxicogenomics Database (CTD).

Results: The identification of three immune subtypes in SSc samples was based on the expression profiles of immune cells. The utilization of 19 key intersectional DEGs among subtypes facilitated the classification of SSc patients into three robust subtypes (gene_ClusterA-C). Gene_ClusterA exhibited significant enrichment of B cells, while gene_ClusterC showed significant enrichment of monocytes. Moderate activation of various immune cells was observed in gene_ClusterB. We identified 8 feature genes. The SVM model demonstrating superior diagnostic performance. Furthermore, correlation analysis revealed a robust association between the feature genes and immune cells. Eight pertinent compounds, namely methotrexate, resveratrol, paclitaxel, trichloroethylene, formaldehyde, silicon dioxide, benzene, and tetrachloroethylene, were identified from the CTD.

Conclusion: The present study has effectively devised an innovative molecular subtyping methodology for patients with SSc and a diagnostic model based on machine learning to aid in clinical treatment. The study has identified potential molecular targets for therapy, thereby offering novel perspectives for the treatment and investigation of SSc.

1 Introduction

Systemic sclerosis (SSc), commonly referred to as scleroderma, is a rare autoimmune disease affecting connective tissues, characterized by skin and internal organ fibrosis, autoimmunity, and vasculopathy, with a significantly higher mortality rate compared to other rheumatic diseases. The progression rate, disease manifestations, and response to therapy exhibit significant variability among individuals (1, 2). Due to its low prevalence, SSc is considered an orphan disease, and its burden is substantial (3). The exact cause of SSc remains uncertain, although considerable evidence suggests that genetic and environmental factors significantly contribute to its development (4, 5). While Raynaud's phenomenon and fatigue are common early symptoms of SSc, their presentation can vary, making it challenging for clinicians to accurately diagnose the disease (6). This diagnostic difficulty may have implications for treatment decisions and patient outcomes.

Presently, the management of SSc centers on addressing the symptoms of affected cutaneous and internal organs, including but not limited to pulmonary, renal, cardiac, pulmonary arterial hypertension, gastrointestinal, and musculoskeletal involvement (2). Conventional therapeutic approaches encompass pharmacological interventions such as cyclophosphamide (CYC) and mycophenolate mofetil (MMF), while hematopoietic stem cell transplantation (HSCT) represents a crucial treatment modality. Recent research has investigated novel pharmacological interventions for the management of SSc, such as rituximab and tocilizumab, among others. The principal immunological indicators and therapeutic objectives implicated in the pathogenesis of SSc have been identified, including IL-6, IL-4, IL-13, TGF-B, and others (7). While significant advancements have been achieved in the investigation and clinical management of SSc pathogenesis, further comprehensive inquiry remains necessary.

In the early stages of SSc, the primary event is vascular injury, which triggers endothelial activation, inflammation mediated by both innate and adaptive immune responses, vascular remodeling, and ultimately fibrosis (2, 8). As such, an examination of the gene expression profiles of peripheral blood mononuclear cells (PBMCs) in SSc patients is of particular significance in comprehending the pathogenesis, immune characteristics, subtyping, and clinical management of SSc patients. At present, SSc is typically categorized into subtypes according to the degree of skin involvement, namely diffuse cutaneous SSC and limited cutaneous SSC (9). This classification based on skin involvement holds significant clinical implications (1). Additionally, a minor subset of SSc patients, known as sine scleroderma, exhibit no skin involvement (10). Presently, there are no alternative or superior subtype definitions that can effectively guide the clinical management of SSc, which poses significant challenges in its treatment.

This study involved the collection of peripheral blood transcriptome datasets from five SSc gene expression datasets sourced from the Gene Expression Omnibus (GEO) database. Through the use of unsupervised machine learning methods, three distinct and reliable subtypes of SSc patients were identified. The exploration of the immune and molecular characteristics of these subtypes has yielded significant insights that are relevant to the advancement of research and treatment of SSc. Moreover, a machine learning diagnostic model was developed utilizing key genes to aid in the clinical management of SSc. This investigation considers the vascular alterations that occur during the initial phases of SSc and introduces an innovative and presently limited technique for characterizing SSc subtypes, providing a fresh outlook for the clinical diagnosis and management of SSc.

2 Materials and methods

2.1 Data acquisition

Peripheral blood gene expression data of SSc patients were collected from the GEO database, encompassing five datasets: GSE130953 (11), GSE22356 (12), GSE65336 (13), GSE33463 (14), and GSE179153 (15). The baseline data of the patients were extracted from the datasets. The GSE179153 dataset was utilized for constructing the machine learning diagnostic model, while the remaining datasets were employed for analysis. The analysis involved 120 SSc patient samples and 113 healthy donor samples. The microarray datasets were obtained from Affymetrix. The raw "CEL" files were acquired and subjected to background adjustment and quantile normalization to produce gene expression matrix files. The probe annotation of the expression matrix was conducted using the R 'idmap2' package. The correlation between patient samples within the analysis dataset was calculated utilizing the R base function 'cor', and samples with a correlation coefficient below 0.7 were eliminated. The batch effects between datasets were eliminated using the "ComBat" algorithm from the 'sva' package. Subsequently, a total of 120 samples from patients with SSc were utilized for analysis. The datasets employed in this study have been succinctly outlined in Table 1.

2.2 Immune infiltration analysis

The "xCell" package, a tool that is presently accessible for identifying cell types across various data sources (16), was utilized in our study to evaluate immune cell infiltration in SSc samples. We employed 64 cell types to characterize the peripheral blood immune cell populations of SSc patients and computed the peripheral blood immune scores.

2.3 Unsupervised consensus clustering in SSc

The identification of intrinsic subgroups with shared biological features can be achieved through the utilization of the "ConsensusClusterPlus" software package in R (17). In order to investigate potential subtypes of SSc patients, we employed the

TABLE 1 The dataset used in study.

Data set	Subjects	Experiment type	Platforms	Tissue
GSE179153	49 SSc vs 25 HC	Expression profiling by array	GPL10558	Whole Blood
GSE130953	62 SSc vs 62 HC	Expression profiling by array	GPL10558	Whole Blood
GSE22356	10 SSc vs 10 HC	Expression profiling by array	GPL570	PBMC
GSE33463	19 SSc vs 41 HC	Expression profiling by array	GPL6947	PBMC
GSE65336	29SSc	Expression profiling by array	GPL570	Whole Blood

"ConsensusClusterPlus" package for unsupervised clustering. The K-Means algorithm based on Euclidean distance and Ward-D linkage was utilized in the analysis, with 1000 iterations performed to ensure classification stability. The cumulative distribution function (CDF) values and the incremental area under the CDF curve were employed as evaluation criteria for each cluster in the consensus clustering process. Subsequently, the clustering results were validated through the utilization of principal component analysis (PCA).

2.4 Identification of differentially expressed genes between subtypes

The Limma package was employed to discern dissimilarly expressed genes among subtypes, utilizing the false discovery rate (FDR) technique to regulate false positives. Significance was established at adjusted p-values of <= 0.05, while a fold change of >= 0.32 was deemed indicative of significant differences.

2.5 Characterization of SSc subtypes

The present study employed the "xCell" package to assess the enrichment of 64 cell types and calculate immune scores in the robust SSc subtypes, in order to characterize them. Additionally, SSc-related immune pathways were selected from published literature and gene set enrichment analysis (GSEA) results, utilizing gene sets derived from the KEGG and Reactome databases, to evaluate the enrichment of metabolic pathways among SSc patient subtypes. Furthermore, the Wilcoxon test was employed to evaluate the enrichment scores of distinct cell types and pathway activities across the three subtypes, where statistical significance was determined at a p-value threshold of less than 0.05.

2.6 Construction of machine learning diagnostic models

Feature genes were selected using the Least Absolute Shrinkage and Selection Operator (LASSO) algorithm, Support Vector Machine (SVM) algorithm, Random forest (RF) and Boruta, based on the intersection of 19 genes among the three robust subtypes. LASSO is a well-established algorithm in machine learning that is commonly employed for feature selection and

data dimensionality reduction. SVM is a supervised machine learning algorithm that can effectively classify high-dimensional large data into a limited number of data points (support vectors), thereby achieving dimensionality reduction. The R packages "glmnet" and "e1071" were utilized to implement LASSO and SVM, respectively. The RF algorithm, which comprises multiple decision trees, was implemented using the R package "randomForest". The Boruta algorithm is a feature selection method used to identify important features in a dataset that have statistical significance. It is implemented using the R package "Boruta". Following the acquisition of feature genes, an assessment of the correlation between these genes and immune cells was conducted. Subsequently, the dataset was partitioned into training and validation sets in a 7:3 ratio. Diagnostic models were constructed using the SVM, RF, and Logistic Regression (LR) algorithms. The LR model, which is a generalized linear regression analysis model, is frequently employed in data mining and disease diagnosis. Its implementation is carried out through the utilization of the base function "glm" in the R programming language. Ultimately, the efficacy of the three models was evaluated in both the training and testing sets by means of Receiver Operating Characteristic (ROC) curves.

2.7 Identification of compounds associated with SSc

The Comparative Toxicogenomics Database (CTD) was utilized to conduct a search for SSc, with a subsequent filtration of compounds associated with the feature genes, as per the "Chemical-Gene Interactions" tab.

3 Results

3.1 Exploring subtypes in SSc

The present study conducted an initial investigation into subtypes of SSc by analyzing peripheral blood expression profiles from a total of 120 SSc patients across four cohorts. To mitigate batch effects between datasets, the ComBat algorithm was employed, and the resulting batch effect-corrected changes were visualized using PCA (Figures 1A, B). Additionally, the xCell package was utilized to perform convolution on the peripheral blood expression profiles of the 120 SSc patients. Based on the
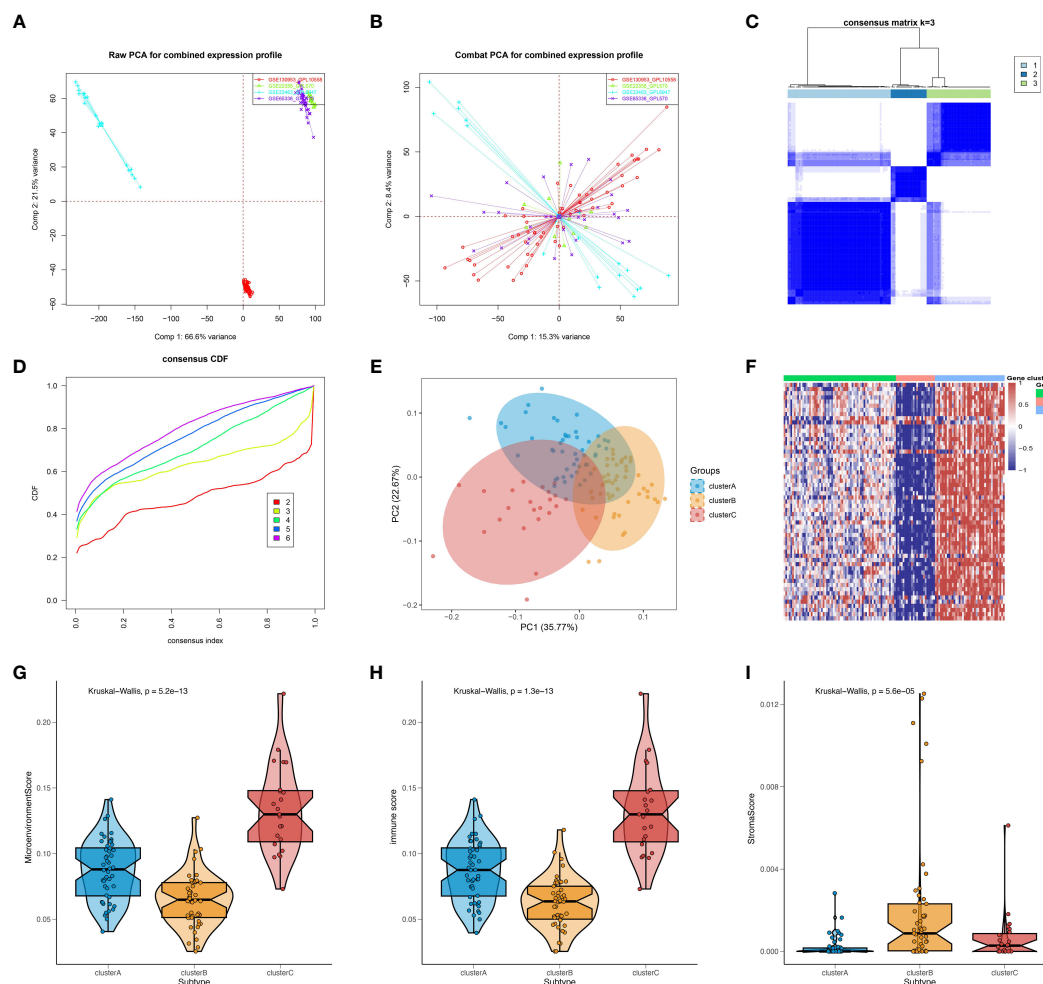
FIGURE 1

Preliminary investigation and molecular features of SSc subtypes. PCA of expression matrix for four different datasets before batch correction **(A)** and after batch correction **(B)**. **(C)** Heatmap of consensus matrix at k = 3. **(D)** Cumulative distribution frequency (CDF) curve of clustered samples. **(E)** PCA plot showing three subtypes after classification. **(F)** Heatmap showing gene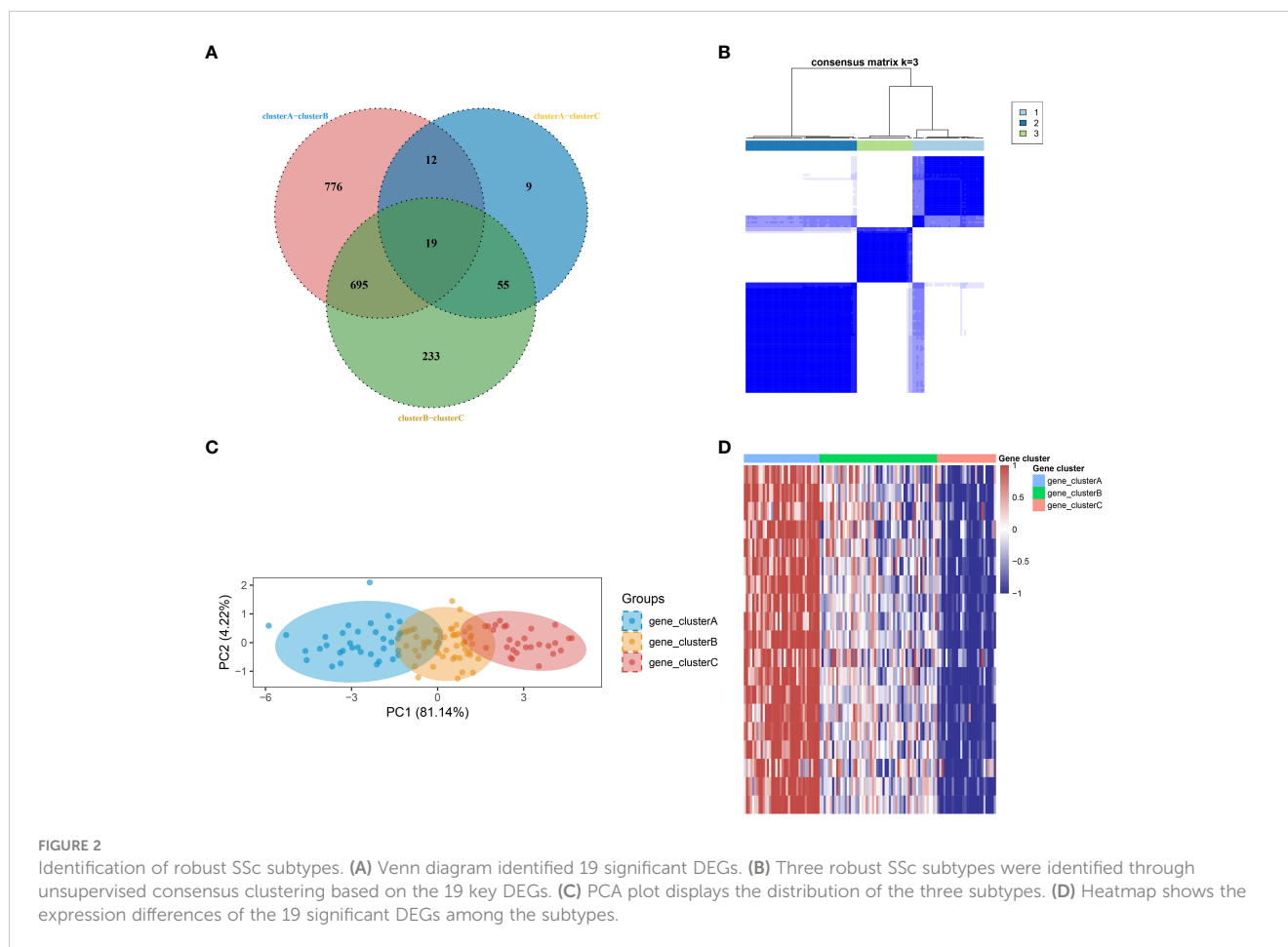 expression differences between subtypes. Microenvironment scores **(G)**, immune scores **(H)**, and stromal scores **(I)** for different subtypes were calculated using the Xcell package.

observed differences in immune cells, the km algorithm with 1000 iterations from the ConsensusClusterPlus package was employed to perform clustering. Through an analysis of the CDF values and the incremental area under the CDF curve, we arrived at the determination that k=3 represents the optimal number of clusters, a finding that was subsequently confirmed by PCA (Figures 1C–E). To further investigate the differences in gene expression between the identified subtypes, we generated a heatmap, which revealed a significant upregulation of genes in cluster C, a significant downregulation of genes in cluster B, and intermediate expression levels in cluster A (Figure 1F). Moreover, we computed microenvironment scores, immune scores, and stromal scores for the various subtypes. The results demonstrated that cluster C had the highest microenvironment and immune scores, cluster B had the lowest scores, and cluster A fell between clusters B and C (Figures 1G, H), which corresponded to the heatmap results. As for the stromal score, cluster B had the highest score, cluster A had the lowest score, and Cluster C was in the middle (Figure 1I). Since the

gene expression data were derived from peripheral blood, the stromal score might not be meaningful. However, overall, these results suggest that stratifying SSc patients based on the immune cell composition in peripheral blood is effective. In summary, the findings indicate that the stratification of SSC patients according to the composition of immune cells in the peripheral blood is a viable approach.

3.2 Identification of robust subtypes in SSc

In order to develop a more comprehensive definition of SSc subtypes, the limma package was utilized to compute differential gene expression among the three subtypes. Through the implementation of a Venn diagram, 19 significant DEGs were identified (Figure 2A), which served as crucial factors in distinguishing the three subtypes. Subsequently, an unsupervised clustering analysis was conducted on the SSc samples, resulting in

FIGURE 2
Identification of robust SSc subtypes. **(A)** Venn diagram identified 19 significant DEGs. **(B)** Three robust SSc subtypes were identified through unsupervised consensus clustering based on the 19 key DEGs. **(C)** PCA plot displays the distribution of the three subtypes. **(D)** Heatmap shows the expression differences of the 19 significant DEGs among the subtypes.

the identification of three more robust subtypes (gene_clusterA-C) using the ConsensusClusterPlus algorithm (Figure 2B), based on the aforementioned 19 key DEGs. The PCA further confirmed the findings (Figure 2C). The heatmap revealed that gene_clusterA manifested significantly elevated expression levels across all 19 genes, gene_clusterC exhibited significantly reduced expression levels, and gene_clusterB displayed moderate expression levels (Figure 2D).

3.3 Molecular features of robust subtypes in SSc

In order to comprehend the molecular attributes and physiological roles of the three resilient subtypes, we conducted an investigation into their prevalence across 64 cell types and immune-related pathways (Figure 3). Our findings indicate that gene_clusterA subtype demonstrated a notably greater prevalence of B cells and T cells in comparison to the other subtypes, particularly in B cell-related enrichments and the B cell receptor signaling pathway. Furthermore, gene_clusterA exhibited a high degree of enrichment in TCR signaling transduction and CD28-dependent PI3K-AKT signaling. Monocytes exhibited anomalous activity in gene_clusterC, which was notably enriched in interleukin-related responses, encompassing
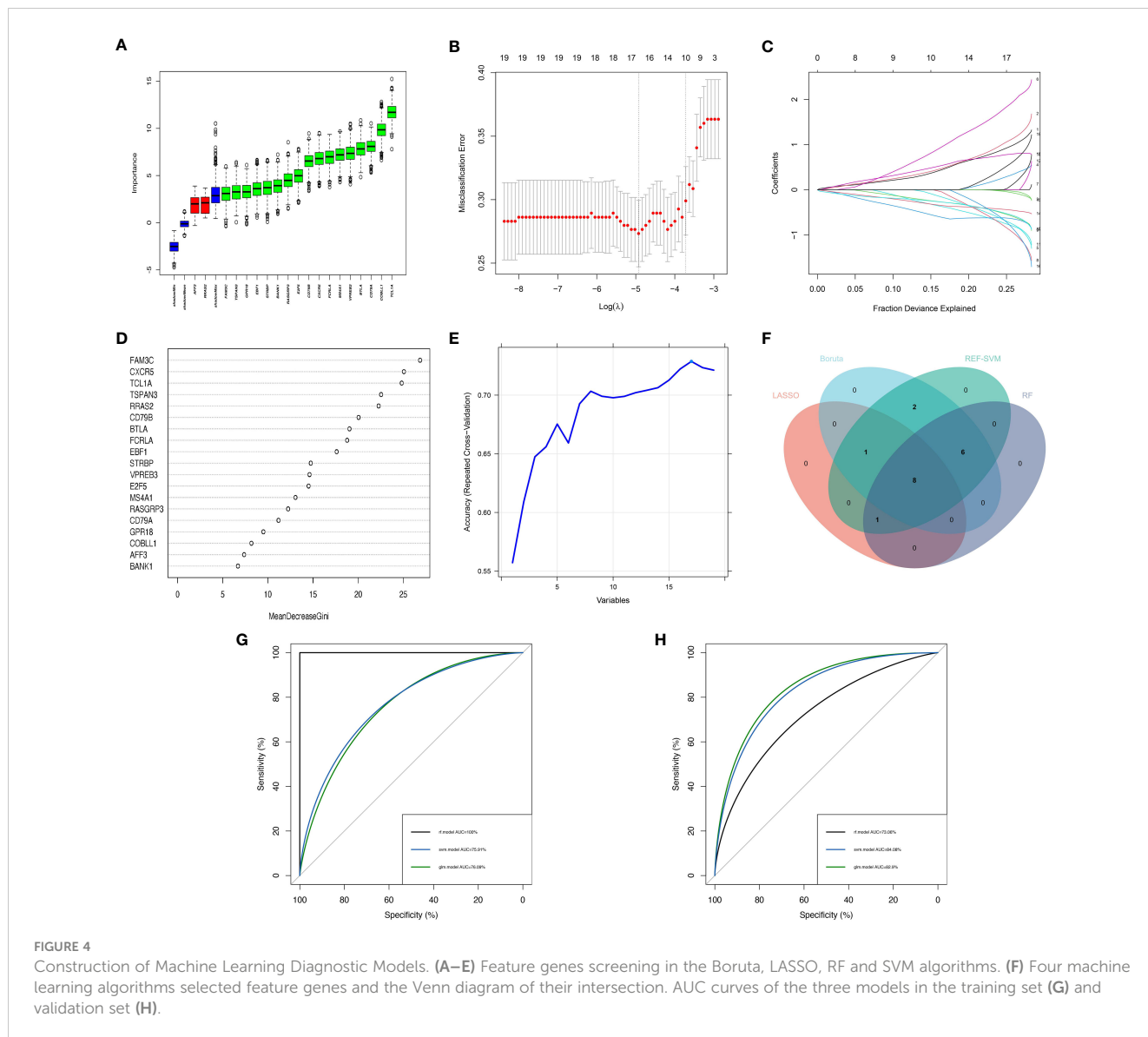
interleukin 1, 6, 10, and 17 signalings, as well as the processing of interleukin 1. Additionally, gene_clusterC demonstrated elevated scores in various signaling pathways, including the chemokine signaling pathway, cytokine-cytokine receptor interaction, mTOR signaling pathway, Nod-like receptor signaling pathway, Notch signaling pathway, PPAR signaling pathway, Toll-like receptor signaling pathway, and VEGF signaling pathway. It is noteworthy that gene_clusterB demonstrates moderate activation throughout all cells and pathways. Utilizing these molecular characteristics, we have classified the gene_clusterA subtype as B-cell rich, gene_clusterB as intermediate, and gene_clusterC as monocyte activated.

3.4 Construction of machine learning diagnostic models

In order to enhance the precision of marker genes, we utilized SVM, LASSO regression, RF and Boruta for feature selection from a pool of 19 key DEGs (Figures 4A–E). Ultimately, we identified 8 feature genes, namely "FAM3C", "BTLA", "STRBP", "RASGRP3", "CD79A", "MS4A1", "CXCR5", and "TCL1A" (Figure 4F). To establish dependable clinical classifiers for SSc subtypes, we developed classification models using SVM, RF, and LR. The dataset GSE179153 was partitioned into a training set and a validation set at

FIGURE 3

Molecular features of subtypes exhibited by immune cells and immune-related pathways. **(A)** Enrichment scores of immune cell infiltration in different subtypes. **(B)** Enrichment scores of SSc-related immune pathways in different subtypes.

a ratio of 7:3. The efficacy of the SVM, RF, and LR models was evaluated in both sets using ROC curves. In the training set, the AUC values of the SVM, RF, and LR models were 0.7591, 1.000, and 0.7609, respectively (Figure 4G). In the validation set, the AUC values were 0.8408, 0.7306, and 0.829, respectively (Figure 4H). Overall, the three machine learning models, which were based on the 8 feature genes, demonstrated exceptional predictive performance, with the SVM model exhibiting the highest performance.

3.5 Correlation between feature genes and immune cells

In order to examine the association between the diagnostic model and SSc, an analysis was conducted to determine the correlation between the 8 feature genes and immune cells present in the peripheral blood of SSc patients (Figure 5). Notably, these 8 genes exhibited a robust correlation with B cells (including naïve B

FIGURE 4

Construction of Machine Learning Diagnostic Models. **(A–E)** Feature genes screening in the Boruta, LASSO, RF and SVM algorithms. **(F)** Four machine learning algorithms selected feature genes and the Venn diagram of their intersection. AUC curves of the three models in the training set **(G)** and validation set **(H)**.

cells, memory B cells, and other B cell subtypes), monocytes, and epithelial cells. This discovery is in strong agreement with our prior analysis and serves to reinforce the precision of our SSc subtype classification.

3.6 Exploration of potential therapeutic drugs for SSc

Nine compounds of relevance were screened in the CTD, based on the 8 feature genes (Table 2). The expression of BTLA is decreased by trichloroethylene, while methotrexate affects the expression of CD79A and formaldehyde results in a decrease in CD79A expression. CXCR5 expression is increased by silicon dioxide. Additionally, the sensitivity of SSc patients to paclitaxel is influenced by the FAM3C protein. Silicon dioxide has been observed to elicit an upregulation of MS4A1 expression, whereas the administration of [hydroxychloroquine + methotrexate +

sulfasalazine] combination therapy has been shown to induce a downregulation of MS4A1 expression. Conversely, formaldehyde exposure has been associated with a reduction in MS4A1 expression. Analogues of silicon dioxide or [rheumatoid arthritis drugs combined with methotrexate] have both been found to result in a decrease in RASGRP3 expression. Resveratrol and benzene exposure have been linked to a decrease in STRBP expression, while trichloroethylene and tetrachloroethylene exposure have been associated with an increase in STRBP gene expression. Additional investigation is necessary to examine the correlation between said compounds and SSc.

4 Discussion

SSc is a chronic fibrotic disease that arises from autoimmune dysfunction (18). It poses a rare and formidable challenge for treatment (19, 20). Early vascular damage in the disease

FIGURE 5
Correlation between the 8 feature genes and immune cell.
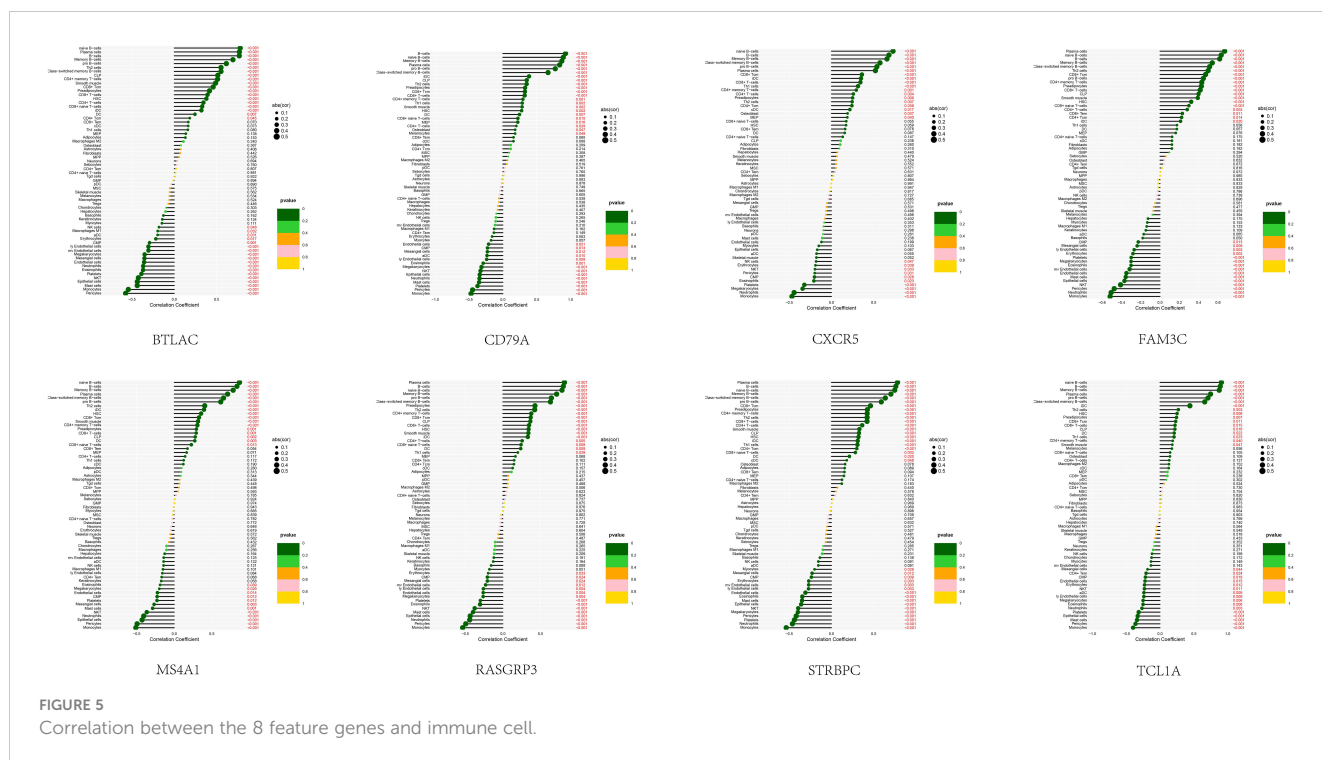
progression serves as a link between immune abnormalities and fibrosis, thereby triggering pathological cascades in multiple organs (21, 22). Furthermore, the early detection of SSc onset is arduous, and prior research has not established a definitive molecular subtype classification for SSc, and there is a dearth of valuable biomarkers for the disease. This study has successfully identified

three distinct and resilient subtypes of SSc, specifically characterized as B-cell rich, intermediate, and monocytes activate types. Subsequently, a diagnostic model has been developed to aid in clinical management.

The xCell algorithm was utilized in this study to examine peripheral blood expression profiles from four datasets of SSc

TABLE 2 Interactions between compounds and genes related to SSc in the CTD database.

Chemical Name	Gene Symbol	Interaction
Trichloroethylene	BTLA	Trichloroethylene results in decreased expression of BTLA mRNA
Methotrexate	CD79A	Methotrexate affects the expression of CD79A mRNA
Trichloroethylene	CD79A	Trichloroethylene results in increased expression of CD79A mRNA
Formaldehyde	CD79A	Formaldehyde results in decreased expression of CD79A mRNA
Silicon Dioxide	CXCR5	Silicon Dioxide results in increased expression of CXCR5 mRNA
Paclitaxel	FAM3C	FAM3C protein affects the susceptibility to Paclitaxel
Silicon Dioxide	MS4A1	Silicon Dioxide results in increased expression of MS4A1 mRNA
Methotrexate	MS4A1	[Hydroxychloroquine co-treated with Methotrexate co-treated with Sulfasalazine] results in decreased expression of MS4A1 mRNA
Formaldehyde	MS4A1	Formaldehyde results in decreased expression of MS4A1 mRNA
Silicon Dioxide	RASGRP3	Silicon Dioxide analog results in decreased expression of RASGRP3 mRNA
Antirheumatic	RASGRP3	[Antirheumatic Agents co-treated with Methotrexate] results in decreased expression of RASGRP3 mRNA
Resveratrol	STRBP	Resveratrol results in decreased expression of STRBP protein
Benzene	STRBP	Benzene results in decreased expression of STRBP mRNA
Trichloroethylene	STRBP	Trichloroethylene results in increased expression of STRBP mRNA
Tetrachloroethylene	STRBP	Tetrachloroethylene results in increased expression of STRBP mRNA

patients. Our findings revealed notable variations in immune cells among SSc patients, which were used to conduct preliminary clustering. Further differential analysis was performed between the identified clusters.By employing the 19 essential DEGs, we have successfully distinguished three resilient subtypes of SSc patients, namely gene_clusterA (characterized by B cell enrichment), gene_clusterB (intermediate in nature), and gene_clusterC (marked by monocyte activation). Significantly, our results were validated by prior investigations (23, 24). Patients with SSc who exhibited a subtype enriched with B cells demonstrated heightened activity of diverse B cell subpopulations in their peripheral blood, including memory and naïve B cells. Notably, B cells play a pivotal role in the pathogenesis of SSc by producing cytokines such as IL-6 and TGF-β (25, 26), engaging in self-activation with T cells (27), stimulating fibroblasts (28), and contributing to endothelial cell activation and injury (29, 30), among other pathways, which ultimately lead to the inflammatory and fibrotic phenotypic manifestations of SSc. Moreover, the B cell-enriched subtype of SSc patients exhibited a significant enrichment in two signaling pathways, namely the CD28-dependent PI3K-AKT signaling and TCR signaling. The activation of PI3K by the co-stimulatory receptor CD28 leads to the generation of PIP3 on the plasma membrane. Akt is involved in the CD28-mediated co-stimulation of T cell activation (31, 32). Furthermore, individuals with SSc who exhibit a monocyte-activated subtype demonstrate a noteworthy increase in the abundance of monocytes in their peripheral blood. This observation is consistent with prior research, such as the work of Alain Lescoat et al, which posits that monocyte adhesion may escalate in SSc due to the loss of CD52 (33). Macrophages derived from monocytes expressing CD163 or CD204 may serve as potential regulators of fibrosis in the skin of individuals with SSc (34, 35). The utilization of flow cytometry by Laure Ricard et al. revealed a noteworthy elevation in 6-Sulfo LacNAc monocytes, intermediate monocytes, and non-classical monocytes in individuals with SSc (24), with a more pronounced increase in SlanMo cells observed in those with diffuse SSc. Furthermore, the subtype activated by monocytes exhibited a notable enrichment in interleukin-mediated responses, encompassing signaling pathways for interleukin-1, interleukin-6, interleukin-10, interleukin-17, and interleukin-1 processing. Interleukins are recognized as significant contributors to the advancement of SSc (36–39). In conclusion, our identification and description of SSc subtypes may serve as promising avenues for future therapeutic research in SSc.

Dimensionality reduction was performed on the 19 key DEGs using SVM, RF, Boruta and LASSO regression, resulting in the identification of 8 feature genes, namely FAM3C, BTLA, STRBP, RASGRP3, CD79A, MS4A1, CXCR5, and TCL1A. Subsequently, clinical diagnostic models were constructed based on the aforementioned 8 feature genes. The model shows good predictive performance in both training set and validation set. BTLA, a constituent of the CD28 superfamily, plays a pivotal role as a co-signaling molecule. Its principal role involves hindering the activation and proliferation of T cells, B cells, and DC cells. Recent investigations have shed light on the notable importance

of BTLA in the realm of autoimmune diseases, as it has demonstrated efficacy in mitigating conditions such as multiple sclerosis (MS), active systemic lupus erythematosus, and rheumatoid arthritis (RA) (40). STRBP is a protein that exhibits affinity for nuclear RNA in spermatids. Trang T Le et al. employed machine learning algorithms to identify a potential association between STRBP and the differentiation cluster cell surface biomarker in the blood of patients with SLE (41). CD79, consisting of CD79A and CD79B, is predominantly expressed in B cells and B-cell tumors, and plays a crucial role in the expression and function of B-cell antigen receptors. CD79A can be utilized as a primary diagnostic marker for B-cell-related diseases (42), and has been implicated in various pathological conditions. Notably, Ian R Hardy and colleagues have proposed the use of monoclonal antibodies targeting CD79B as a means to collectively suppress B cells and prevent autoimmunity, with the added benefit of facilitating rapid immune recovery, unlike other approaches that induce B cell death (43). This implies that targeting CD79A has significant potential for the treatment of SSc. The gene MS4A1, also known as CD20, encodes a surface molecule present on B-cells that plays a crucial role in their development and differentiation into plasma cells. It is worth noting that rituximab, a chimeric antibody specifically targeting CD20, has exhibited effectiveness in treating fibrotic lesions in SSc and has been approved for the management of SSc and SSc-ILD in certain countries (44). The CXCR5 receptor interacts with CXCL13, a chemoattractant that attracts B-cells. The CXCL13-CXCR5 axis fulfills various biological functions, including the regulation of cancer cell growth, proliferation, invasion, and metastasis (45). Additionally, this axis is implicated in the pathogenesis of several autoimmune diseases (46). TCL1A functions as a co-activator of the serine/threonine kinase Akt, facilitating cell survival, growth, and proliferation through various interactions. TCL1A has the ability to modulate B-cell differentiation and regulation (47). In summary, these feature genes are highly associated with immunity, which provides an important reference for exploring their role in SSc.

By utilizing the CTD, we have successfully identified 9 compounds that exhibit an association with SSc and exert an impact on the eight feature genes. Notably, methotrexate stands out as the most frequently employed immunosuppressant in SSc patients (48). The European League Against Rheumatism advocates for methotrexate as the primary treatment option for early diffuse SSc's skin manifestations (49). The impact of Resveratrol on SSc has been noted, as it has the potential to enhance fibrosis and mitigate inflammatory responses in SSc through the modulation of the SIRT1/mTOR signaling pathway (50). Paclitaxel is a highly efficacious natural anticancer agent; however, there have been documented cases of SSc development in cancer patients undergoing paclitaxel treatment (51). The precise mechanisms underlying this phenomenon remain to be elucidated. Additionally, the other compounds, namely trichloroethylene, formaldehyde, benzene, and tetrachloroethylene, are all recognized environmental exposure factors associated with SSc. These compounds are typical occupational exposure substances, and their association with SSc has been reported in the scientific literature (52–54).

5 Conclusion

This study utilized peripheral blood expression profiling data to identify three molecular subtypes of SSc and examined their molecular characteristics. Additionally, a machine learning diagnostic model was developed to aid in clinical identification. Furthermore, this investigation revealed previously unexplored therapeutic targets and compounds for SSc, offering novel insights for future research in this field. Despite the utilization of rigorous bioinformatics methods, this study is not without limitations. It is imperative to conduct molecular experimental validation to corroborate the findings. Furthermore, additional comprehensive research is required to fully explore the potential therapeutic targets and related drugs for SSc.

Data availability statement

The datasets supporting the conclusions of this article are available in the GEO repository, (https://www.ncbi.nlm.nih.gov/geo/). The accession number(s) can be found in the article.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the patients/participants or patients/participants legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

Author contributions

QW: Methodology, Project administration, Software, Supervision, Writing – original draft, Writing – review & editing. C-LL: Writing – original draft, Writing – review & editing. LW: Writing – review & editing. J-YH: Writing – review & editing. QY: Writing – review & editing. S-XZ: Writing – review & editing. P-FH: Writing – review & editing.

Funding

Acknowledgments

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Denton CP, Khanna D. Systemic sclerosis. *Lancet* (2017) 390:1685–99. doi: 10.1016/S0140-6736(17)30933-9

2. Volkmann ER, Andréasson K, Smith V. Systemic sclerosis. *Lancet* (2023) 401:304–18. doi: 10.1016/S0140-6736(22)01692-0

3. Morrisroe K, Stevens W, Sahhar J, Ngian GS, Ferdowsi N, Hansen D, et al. The clinical and economic burden of systemic sclerosis related interstitial lung disease. *Rheumatol (Oxford)* (2020) 59:1878–88. doi: 10.1093/rheumatology/kez532

4. Bossini-Castillo L, López-Isac E, Mayes MD, Martín J. Genetics of systemic sclerosis. *Semin Immunopathol* (2015) 37:443–51. doi: 10.1007/s00281-015-0499-z

5. Ferri C, Arcangeletti MC, Caselli E, Zakrzewska K, Maccari C, Calderaro A, et al. Insights into the knowledge of complex diseases: Environmental infectious/toxic agents as potential etiopathogenetic factors of systemic sclerosis. *J Autoimmun* (2021) 124:102727. doi: 10.1016/j.jaut.2021.102727

6. Assassi S, Leyva AL, Mayes MD, Sharif R, Nair DK, Fischbach M, et al. Predictors of fatigue severity in early systemic sclerosis: a prospective longitudinal study of the GENISOS cohort. *PloS One* (2011) 6:e26061. doi: 10.1371/journal.pone.0026061

7. Bukiri H, Volkmann ER. Current advances in the treatment of systemic sclerosis. *Curr Opin Pharmacol* (2022) 64:102211. doi: 10.1016/j.coph.2022.102211

8. Truchetet ME, Brembilla NC, Chizzolini C. Current concepts on the pathogenesis of systemic sclerosis. *Clin Rev Allergy Immunol* (2023) 64:262–83. doi: 10.1007/s12016-021-08889-8

9. Hughes M, Herrick AL. Systemic sclerosis. *Br J Hosp Med (Lond)* (2019) 80:530–6. doi: 10.12968/hmed.2019.80.9.530

10. Diab S, Dostrovsky N, Hudson M, Tatibouet S, Fritzler MJ, Baron M, et al. Systemic sclerosis sine scleroderma: a multicenter study of 1417 subjects. *J Rheumatol* (2014) 41:2179–85. doi: 10.3899/jrheum.140236

11. Assassi S, Wang X, Chen G, Goldmuntz E, Keyes-Elstein L, Ying J, et al. Myeloablation followed by autologous stem cell transplantation normalises systemic sclerosis molecular signatures. *Ann Rheum Dis* (2019) 78:1371–8. doi: 10.1136/annrheumdis-2019-215770

12. Risbano MG, Meadows CA, Coldren CD, Jenkins TJ, Edwards MG, Collier D, et al. Altered immune phenotype in peripheral blood cells of patients with scleroderma-

associated pulmonary hypertension. *Clin Transl Sci* (2010) 3:210–8. doi: 10.1111/j.1752-8062.2010.00218.x

13. Guo X, Higgs BW, Bay-Jensen AC, Karsdal MA, Yao Y, Roskos LK, et al. Suppression of T cell activation and collagen accumulation by an anti-IFNAR1 mAb, anifrolumab, in adult patients with systemic sclerosis. *J Invest Dermatol* (2015) 135:2402–9. doi: 10.1038/jid.2015.188

14. Cheadle C, Berger AE, Mathai SC, Grigoryev DN, Watkins TN, Sugawara Y, et al. Erythroid-specific transcriptional changes in PBMCs from pulmonary hypertension patients. *PLoS One* (2012) 7:e34951. doi: 10.1371/journal.pone.0034951

15. Farutin V, Kurtagic E, Pradines JR, Capila I, Mayes MD, Wu M, et al. Multiomic study of skin, peripheral blood, and serum: is serum proteome a reflection of disease process at the end-organ level in systemic sclerosis? *Arthritis Res Ther* (2021) 23:259. doi: 10.1186/s13075-021-02633-5

16. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol* (2017) 18:220. doi: 10.1186/s13059-017-1349-1

17. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* (2010) 26:1572–3. doi: 10.1093/bioinformatics/btq170

18. Rosendahl AH, Schönborn K, Krieg T. Pathophysiology of systemic sclerosis (scleroderma). *Kaohsiung J Med Sci* (2022) 38:187–95. doi: 10.1002/kjm2.12505

19. Chifflot H, Fautrel B, Sordet C, Chatelus E, Sibilia J. Incidence and prevalence of systemic sclerosis: a systematic literature review. *Semin Arthritis Rheum* (2008) 37:223–35. doi: 10.1016/j.semarthrit.2007.05.003

20. Tyndall AJ, Bannert B, Vonk M, Airò P, Cozzi F, Carreira PE, et al. Causes and risk factors for death in systemic sclerosis: a study from the EULAR Scleroderma Trials and Research (EUSTAR) database. *Ann Rheum Dis* (2010) 69:1809–15. doi: 10.1136/ard.2009.114264

21. Asano Y. The pathogenesis of systemic sclerosis: an understanding based on a common pathologic cascade across multiple organs and additional organ-specific pathologies. *J Clin Med* (2020) 9(9):2687. doi: 10.3390/jcm9092687

22. Avouac J, Fransen J, Walker UA, Riccieri V, Smith V, Muller C, et al. Preliminary criteria for the very early diagnosis of systemic sclerosis: results of a Delphi Consensus Study from EULAR Scleroderma Trials and Research Group. *Ann Rheum Dis* (2011) 70:476–81. doi: 10.1136/ard.2010.136929

23. Humby F, Durez P, Buch MH, Lewis MJ, Rizvi H, Rivellese F, et al. Rituximab versus tocilizumab in anti-TNF inadequate responder patients with rheumatoid arthritis (R4RA): 16-week outcomes of a stratified, biopsy-driven, multicentre, open-label, phase 4 randomised controlled trial. *Lancet* (2021) 397:305–17. doi: 10.1016/S0140-6736(20)32341-2

24. Ricard L, Eshagh D, Siblany L, de Vassoigne F, Malard F, Laurent C, et al. 6-Sulfo LacNAc monocytes are quantitatively and functionally disturbed in systemic sclerosis patients. *Clin Exp Immunol* (2022) 209:175–81. doi: 10.1093/cei/uxac059

25. Fielding CA, Jones GW, McLoughlin RM, McLeod L, Hammond VJ, Uceda J, et al. Interleukin-6 signaling drives fibrosis in unresolved inflammation. *Immunity* (2014) 40:40–50. doi: 10.1016/j.immuni.2013.10.022

26. Zi Z, Chapnick DA, Liu X. Dynamics of TGF-β/smad signaling. *FEBS Lett* (2012) 586:1921–8. doi: 10.1016/j.febslet.2012.03.063

27. Bosello S, Angelucci C, Lama G, Alivernini S, Proietti G, Tolusso B, et al. Characterization of inflammatory cell infiltrate of scleroderma skin: B cells and skin score progression. *Arthritis Res Ther* (2018) 20:75. doi: 10.1186/s13075-018-1569-0

28. François A, Chatelus E, Wachsmann D, Sibilia J, Bahram S, Alsaleh G, et al. B lymphocytes and B-cell activating factor promote collagen and profibrotic markers expression by dermal fibroblasts in systemic sclerosis. *Arthritis Res Ther* (2013) 15:R168. doi: 10.1186/ar4352

29. Kill A, Tabeling C, Undeutsch R, Kühl AA, Günther J, Radic M, et al. Autoantibodies to angiotensin and endothelin receptors in systemic sclerosis induce cellular and systemic events associated with disease pathogenesis. *Arthritis Res Ther* (2014) 16:R29. doi: 10.1186/ar4457

30. Sgonc R, Gruschwitz MS, Boeck G, Sepp N, Gruber J, Wick G. Endothelial cell apoptosis in systemic sclerosis is induced by antibody-dependent cell-mediated cytotoxicity via CD95. *Arthritis Rheum* (2000) 43:2550–62. doi: 10.1002/1529-0131(200011)43:11<2550::AID-ANR24>3.0.CO;2-H

31. Okkenhaug K, Bilancio A, Emery JL, Vanhaesebroeck B. Phosphoinositide 3-kinase in T cell activation and survival. *Biochem Soc Trans* (2004) 32:332–5. doi: 10.1042/bst0320332

32. Kane LP, Weiss A. The PI-3 kinase/Akt pathway and T cell activation: pleiotropic pathways downstream of PIP3. *Immunol Rev* (2003) 192:7–20. doi: 10.1034/j.1600-065X.2003.00008.x

33. Lescoat A, Lecureur V, Varga J. Contribution of monocytes and macrophages to the pathogenesis of systemic sclerosis: recent insights and therapeutic implications. *Curr Opin Rheumatol* (2021) 33:463–70. doi: 10.1097/BOR.0000000000000835

34. Higashi-Kuwata N, Jinnin M, Makino T, Fukushima S, Inoue Y, Muchemwa FC, et al. Characterization of monocyte/macrophage subsets in the skin and peripheral blood derived from patients with systemic sclerosis. *Arthritis Res Ther* (2010) 12:R128. doi: 10.1186/ar3066

35. Christmann RB, Lafyatis R. The cytokine language of monocytes and macrophages in systemic sclerosis. *Arthritis Res Ther* (2010) 12:146. doi: 10.1186/ar3167

36. De Luca G, Cavalli G, Campochiaro C, Bruni C, Tomelleri A, Dagna L, et al. Interleukin-1 and systemic sclerosis: getting to the heart of cardiac involvement. *Front Immunol* (2021) 12:653950. doi: 10.3389/fimmu.2021.653950

37. Shima Y. The benefits and prospects of interleukin-6 inhibitor on systemic sclerosis. *Mod Rheumatol* (2019) 29:294–301. doi: 10.1080/14397595.2018.1559909

38. Salim PH, Jobim M, Bredemeier M, Chies JA, Brenol JC, Jobim LF, et al. Interleukin-10 gene promoter and NFKB1 promoter insertion/deletion polymorphisms in systemic sclerosis. *Scand J Immunol* (2013) 77:162–8. doi: 10.1111/sji.12020

39. Wei L, Abraham D, Ong V. The yin and yang of IL-17 in systemic sclerosis. *Front Immunol* (2022) 13:885609. doi: 10.3389/fimmu.2022.885609

40. Ning Z, Liu K, Xiong H. Roles of BTLA in immunity and immune disorders. *Front Immunol* (2021) 12:654960. doi: 10.3389/fimmu.2021.654960

41. Le TT, Blackwood NO, Taroni JN, Fu W, Breitenstein MK. Integrated machine learning pipeline for aberrant biomarker enrichment (i-mAB): characterizing clusters of differentiation within a compendium of systemic lupus erythematosus patients. *AMIA Annu Symp Proc* (2018) 2018:1358–67.

42. Chu PG, Arber DA. CD79: a review. *Appl Immunohistochem Mol Morphol* (2001) 9:97–106. doi: 10.1097/00129039-200106000-00001

43. Hardy IR, Anceriz N, Rousseau F, Seefeldt MB, Hatterer E, Irla M, et al. Anti-CD79 antibody induces B cell anergy that protects against autoimmunity. *J Immunol* (2014) 192:1641–50. doi: 10.4049/jimmunol.1302672

44. Ebata S, Yoshizaki-Ogawa A, Sato S, Yoshizaki A. New era in systemic sclerosis treatment: recently approved therapeutics. *J Clin Med* (2022) 11(15):4631. doi: 10.3390/jcm11154631

45. Hussain M, Adah D, Tariq M, Lu Y, Zhang J, Liu J. CXCL13/CXCR5 signaling axis in cancer. *Life Sci* (2019) 227:175–86. doi: 10.1016/j.lfs.2019.04.053

46. Pan Z, Zhu T, Liu Y, Zhang N. Role of the CXCL13/CXCR5 axis in autoimmune diseases. *Front Immunol* (2022) 13:850998. doi: 10.3389/fimmu.2022.850998

47. Brinas F, Danger R, Brouard S. TCL1A, B cell regulation and tolerance in renal transplantation. *Cells 10* (2021) 10(6):1367. doi: 10.3390/cells10061367

48. Zhu JL, Black SM, Chen HW, Jacobe HT. Emerging treatments for scleroderma/systemic sclerosis. *Fac Rev* (2021) 10:43. doi: 10.12703/r/10-43

49. Kowal-Bielecka O, Fransen J, Avouac J, Becker M, Kulak A, Allanore Y, et al. Update of EULAR recommendations for the treatment of systemic sclerosis. *Ann Rheum Dis* (2017) 76:1327–39. doi: 10.1136/annrheumdis-2016-209909

50. Yao Q, Wu Q, Xu X, Xing Y, Liang J, Lin Q, et al. Resveratrol ameliorates systemic sclerosis via suppression of fibrosis and inflammation through activation of SIRT1/mTOR signaling. *Drug Des Devel Ther* (2020) 14:5337–48. doi: 10.2147/DDDT.S281209

51. Kawakami T, Tsutsumi Y, Soma Y. Limited cutaneous systemic sclerosis induced by paclitaxel in a patient with breast cancer. *Arch Dermatol* (2009) 145:97–8. doi: 10.1001/archdermatol.2008.532

52. Bovenzi M, Barbone F, Pisa FE, Della Vedova A, Betta A, Romeo L, et al. [Scleroderma and occupational risk factors: a case-control study]. *G Ital Med Lav Ergon* (2003) 25 Suppl:46–7.

53. Pralong P, Cavailhes A, Balme B, Cottin V, Skowron F. Diffuse systemic sclerosis after occupational exposure to trichloroethylene and perchloroethylene. *Ann Dermatol Venereol* (2009) 136:713–7. doi: 10.1016/j.annder.2008.10.043

54. Borghini A, Poscia A, Bosello S, Teleman AA, Bocci M, Iodice L, et al. Environmental pollution by benzene and PM(10) and clinical manifestations of systemic sclerosis: A correlation study. *Int J Environ Res Public Health* (2017) 14(11):1297. doi: 10.3390/ijerph14111297