

# Functional States of T Cells in Scleroderma Peripheral Blood Reveal Distinct Cytokine Polarization Axes

Glen Ritschel Affiliation: Ritschel Research, Tega Cay, SC, USA Correspondence: glen.ritschel@gmail.com Running title: T-cell functional states in scleroderma

## Abstract

Systemic sclerosis (scleroderma) is a complex autoimmune disease characterized by immune dysregulation, vasculopathy, and progressive fibrosis. Although T cells are known contributors to disease pathogenesis, their functional organization in peripheral blood remains incompletely defined.

Here, we performed integrative single-cell transcriptomic analysis of peripheral blood mononuclear cells from a large scleroderma cohort using a batch-corrected latent variable model. After stringent quality control and integration, we extracted a high-confidence T-cell compartment and characterized functional states using program-level scoring to recover graded polarization states rather than enforcing discrete helper T-cell clusters.

This approach identified dominant naive/central memory and cytotoxic T-cell populations, along with smaller but transcriptionally coherent Th1, Th2, Th17, T follicular helper, and regulatory T-cell states. Each functional state exhibited a highly specific cytokine and transcription factor signature, including IFNG and TBX21 for Th1 cells, RORC and IL17A for Th17 cells, GATA3 and IL4/IL13 for Th2 cells, BCL6 for T follicular helper cells, and FOXP3 and IL2RA for regulatory T cells. Cytotoxic T cells were characterized by high expression of NKG7, GNLY, and PRF1, while naive and unpolarized cells showed minimal effector cytokine expression.

Together, these findings demonstrate that functional polarization of T cells can be robustly recovered from resting peripheral blood using integrated single-cell transcriptomics and program-based inference. This framework provides a principled approach for interpreting systemic immune activation states in the context of downstream fibrotic pathways in systemic sclerosis.

## Keywords

systemic sclerosis; scleroderma; T cells; single-cell RNA-seq; scVI; cytokines; immune polarization

## Introduction

Systemic sclerosis (SSc) is a heterogeneous autoimmune disease characterized by immune dysregulation, microvascular injury, and progressive fibrosis of the skin and internal organs (1-3,35). Although fibrosis represents the defining clinical manifestation, substantial evidence indicates that immune abnormalities precede overt fibrotic remodeling, implicating early immune-mediated processes in disease initiation and progression (4,5). Both innate and adaptive immune cells contribute to this process through cytokine production, cellular interactions, and modulation of stromal cell behavior, positioning the immune system as a central driver of disease pathogenesis rather than a secondary consequence of tissue injury (6,7,37).

Among adaptive immune populations, T cells have long been implicated in SSc through altered activation states, skewed cytokine profiles, and dysregulated interactions with fibroblasts, endothelial cells, and B cells (8-10,37). Prior studies have reported imbalances in T helper subsets, including Th1, Th2, Th17, regulatory T cells (Tregs), and T follicular helper (Tfh) cells, as well as expansion of cytotoxic T-cell populations (11-14). These subsets have been associated with key pathogenic pathways, such as interferon signaling, IL-17-mediated inflammation, aberrant B-cell activation, and profibrotic cytokine production (15-18). However, the extent to which these functional states are detectable and coherently organized in peripheral blood remains unclear, particularly given the dynamic and plastic nature of T-cell differentiation (19).

Efforts to classify T-cell subsets in SSc peripheral blood using bulk transcriptomics or flow cytometry have provided important insights but are inherently limited by population averaging and restricted marker panels (20-21). Single-cell RNA sequencing (scRNA-seq) offers the potential to resolve immune heterogeneity at higher resolution, yet studies of circulating immune cells in autoimmune disease face substantial technical and biological challenges (22-24). These include strong batch effects across donors and experiments, inter-individual variability in immune composition, and the low abundance of polarized helper T-cell populations in resting peripheral blood mononuclear cells (PBMCs). As a result, discrete clustering of helper T-cell subsets in PBMC scRNA-seq data is often unstable or yields inconsistent subtype definitions (25).

Recent scRNA-seq studies in systemic sclerosis have provided valuable insights into immune dysregulation, particularly in affected tissues such as skin and lung, where effector T-cell populations are more abundant and transcriptionally polarized (27-29). In contrast, analyses of peripheral blood have frequently reported diffuse or overlapping T-cell states, reflecting both biological plasticity and technical limitations. Together, these observations raise a central conceptual question: whether disease-relevant T-cell biology in peripheral blood is best represented by discrete cell types or by graded functional activation states distributed along transcriptional continua.

Advances in probabilistic latent variable models for scRNA-seq integration provide an opportunity to address this challenge (22). Methods such as single-cell variational inference (scVI) enable robust integration of large, multi-sample datasets by modeling technical variation explicitly while preserving biological signal. When combined with program-based inference approaches that score cells for predefined functional gene sets, these methods allow immune states to be characterized without enforcing rigid clustering boundaries (23–25). This framework is particularly well suited to peripheral blood analyses, where T cells often exhibit partial or transient activation rather than stable lineage commitment (26).

In this study, we applied scVI-based integration to a large cohort of scleroderma PBMC samples to characterize T-cell functional organization at single-cell resolution. After stringent quality control and batch correction, we extracted a high-confidence T-cell compartment and inferred functional states using program-level scoring rather than discrete helper T-cell clustering. This approach enabled recovery of coherent Th1, Th2, Th17, Tfh, Treg, cytotoxic, and naive T-cell programs despite their low abundance in circulating blood. By focusing on functional polarization axes rather than categorical cell types, we aimed to bridge systemic immune activation patterns to downstream fibrotic mechanisms implicated in systemic sclerosis.

Together, this work provides a principled framework for interpreting T-cell functional heterogeneity in peripheral blood and establishes a foundation for future studies integrating clinical phenotypes, longitudinal sampling, and tissue-level immune profiling in systemic sclerosis.

## Materials and Methods

### Dataset and preprocessing

Peripheral blood mononuclear cell (PBMC) transcriptomic data were obtained from the public dataset GSE195452. Processed gene-by-cell count matrices were used directly without re-alignment of raw sequencing reads. All samples shared a consistent gene ordering, enabling direct merging across donors. Cells with fewer than 75 detected genes or 75 total counts were removed. Samples with fewer than 25 retained cells after quality control were excluded from downstream analysis.

### Dataset Selection and Justification

We selected GSE195452 as the primary dataset for this analysis based on several key criteria. First, this dataset provides per-cell gene expression matrices suitable for batch-corrected integration using probabilistic latent variable models, rather than only processed embeddings or cluster-level summaries. Second, the dataset contains all major peripheral blood mononuclear cell (PBMC) populations including

B cells, T cells, NK cells, and myeloid cells, enabling comprehensive immune profiling. Third, with over 500 samples after quality control, GSE195452 represents one of the largest publicly available SSc PBMC single-cell RNA sequencing datasets, providing sufficient statistical power to detect rare T cell functional states. Fourth, the consistent gene ordering and matrix format across samples enabled robust technical integration while preserving biological variation. While other recent SSc single-cell studies have been published (30-32), many do not deposit the per-cell expression data required for independent reanalysis and reproducible integration with batch correction methods.

The dataset was generated using standard single-cell RNA sequencing protocols on peripheral blood mononuclear cells isolated from whole blood. Source material consists of PBMCs rather than tissue biopsies, making this dataset appropriate for characterizing circulating immune states rather than tissue-resident populations. This analytical scope aligns with our goal of understanding systemic immune activation patterns that may influence downstream tissue pathology.

#### Analytical Scope and Approach to Metadata Limitations

This analysis focuses on functional state characterization and program-based inference of T cell polarization rather than direct disease-specific differential expression analysis. While the dataset includes samples from both SSc patients and healthy controls at the study design level, incomplete standardization of clinical metadata annotations at the per-cell level in the deposited data precluded robust statistical comparisons between disease and control groups. Specifically, essential metadata fields such as standardized disease status labels, patient demographics, clinical phenotypes, and disease severity measures were not uniformly available across all samples in the public repository.

Given these constraints, we adopted an analytical strategy focused on: (1) characterizing the diversity of T cell functional states present in PBMCs using unsupervised methods, (2) inferring functional programs through gene signature scoring rather than enforcing discrete cluster assignments, and (3) quantifying the per-sample composition of functional states to enable future meta-analyses when integrated with external clinical phenotype data. We present per-sample functional state counts and fractions (Supplementary Tables S3A-B) that can be correlated with clinical variables in subsequent studies with access to complete patient metadata.

This approach is methodologically conservative and ensures that our conclusions are fully supported by the available data without requiring unverifiable assumptions about disease group assignments. The functional state framework we establish is generalizable across disease contexts and provides a foundation for future integrative analyses combining immune cell composition with clinical outcomes.

#### Data integration using scVI

Batch correction and integration were performed using the scVI probabilistic latent variable framework. The model was trained on 3,000 highly variable genes with sample identity as the batch covariate. A 30-dimensional latent space was learned using a negative binomial likelihood, and convergence was assessed by monitoring the evidence lower bound. The resulting latent representation was used for neighborhood graph construction and uniform manifold approximation and projection (UMAP) visualization.

#### T-cell identification and subsetting

T cells were identified using a conservative marker-based scoring approach incorporating CD3D, CD3E, CD247, and TRAC expression. Cells exceeding a high quantile threshold for the T-cell score were retained. Additional scores for B-cell and myeloid markers were used to exclude contaminating populations.

#### Functional state inference

Rather than assigning discrete helper T-cell clusters, functional states were inferred using program-level scoring. Canonical gene sets representing Th1, Th2, Th17, T follicular helper, regulatory T-cell, cytotoxic, and naive programs were evaluated at the single-cell level. Each cell was assigned a dominant functional state if its highest program score exceeded a conservative threshold; cells failing this criterion were labeled as unpolarized.

### Statistical analysis

Statistical analyses were performed in Python 3.10.12 using scanpy 1.9.3, scvi-tools 1.0.0, and standard scientific computing libraries (numpy, pandas, scipy, matplotlib). Differential gene expression was assessed using Wilcoxon rank-sum tests with Benjamini-Hochberg false discovery rate (FDR) correction; adjusted  $p < 0.05$  was considered significant. UMAP visualizations used default scanpy parameters ( $n\_neighbors=15$ ,  $min\_dist=0.1$ ). All analyses used fixed random seeds (42) for reproducibility.

### Results

#### Integrated PBMC analysis reveals structured T-cell organization

Integration of 67,592 PBMCs across 578 samples using scVI produced a well-mixed latent space with preservation of major immune lineages (Fig. 1A,B). Marker-based scoring enabled extraction of a high-confidence T-cell compartment comprising 14,617 cells for downstream analysis (Fig. 1C).

#### Functional T-cell states are recovered using program-based scoring

Graph-based clustering identified a dominant naive/central memory compartment and a distinct cytotoxic population, while helper subsets did not form large discrete clusters (Fig. 2A). Program-based functional inference instead revealed graded polarization states across the latent space, including Th1, Th2, Th17, T follicular helper, and regulatory T cells (Fig. 2B,C).

Program-level inference revealed that approximately half of circulating T cells exhibited a naive or central memory phenotype, while a substantial fraction displayed cytotoxic characteristics. Smaller but transcriptionally coherent Th1, Th2, Th17, T follicular helper, and regulatory T-cell states were also detected. Each functional state demonstrated a highly specific cytokine and transcription factor signature. Th1 cells showed strong IFNG and TBX21 expression, Th17 cells expressed RORC and IL17A, Th2 cells were marked by GATA3 and IL4/IL13, T follicular helper cells by BCL6 and PDCD1, and regulatory T cells by FOXP3 and IL2RA.

#### Cytokine polarization axes distinguish helper T-cell states

Each functional state exhibited a highly specific cytokine and transcription factor signature, with Th1 cells expressing IFNG and TBX21, Th17 cells expressing RORC and IL17A, Th2 cells expressing GATA3 and IL4/IL13, T follicular helper cells expressing BCL6, and regulatory T cells expressing FOXP3 and IL2RA (Fig. 3A,C). IFNG and IL17A expression defined near-orthogonal Th1 and Th17 polarization axes (Fig. 3B). Comparison of cytokine expression across functional states revealed minimal overlap between these axes, with IFN- $\gamma$  expression confined to Th1 and cytotoxic populations and IL-17A expression restricted to a small Th17 subset. These patterns support a model in which helper T-cell activation in peripheral blood exists along graded transcriptional continua rather than as discrete clusters.

Sample-level composition of T-cell functional states was quantified across all samples. Supplementary Table S3A reports absolute counts of T cells assigned to each functional state per sample, while Supplementary Table S3B reports corresponding fractions. Across the cohort, naive/central memory T cells constituted the largest fraction, followed by a substantial cytotoxic compartment. Polarized helper T-cell states, including Th1, Th2, Th17, Tfh, and Treg populations, were consistently detected at lower frequencies across samples. These quantitative summaries establish a cohort-level baseline for future

stratification by clinical phenotype and disease severity when higher-resolution metadata become available.

## Discussion

Using scVI-based integration, we analyzed T cells in a unified latent space while correcting for sample-level batch effects across a large cohort of systemic sclerosis PBMCs. Rather than segregating into large discrete helper clusters, T cells occupied graded transcriptional continua in the latent space, consistent with dynamic polarization states that exist on a continuum rather than as sharp categorical boundaries. Despite their limited abundance in resting peripheral blood, helper T-cell states exhibited near-orthogonal cytokine signatures, supporting the validity of program-based functional inference as an alternative to forced clustering approaches.

The identification of distinct Th2 and Th17 functional states with elevated expression of pro-fibrotic cytokines (IL4, IL13, IL17A) provides a mechanistic link between adaptive immune activation and stromal responses in systemic sclerosis. These T cell-derived cytokines are established activators of fibroblast differentiation and collagen production, with IL-13 and IL-4 directly stimulating myofibroblast differentiation through STAT6 signaling (33), and IL-17A promoting fibroblast activation through NF- $\kappa$ B and MAPK pathways (34,4-6,37). The concurrent presence of activated B cells with enhanced antigen presentation capacity in SSc patients (7) may drive the polarization of these pathogenic T cell states through hyperactive CD80/CD86-CD28 costimulation and elevated MHC class II-mediated antigen presentation (36). This positions T cell functional states as a critical bridge between upstream immune dysregulation and downstream fibrotic remodeling.

The cytotoxic T cell population, characterized by high expression of NKG7, GNLY, PRF1, and GZMB, represents another functionally distinct state with potential relevance to SSc pathogenesis. While classically associated with viral immunity and tumor surveillance, cytotoxic T cells have been implicated in autoimmune tissue damage through direct killing of target cells and production of inflammatory cytokines including IFN- $\gamma$  and TNF- $\alpha$ . The balance between regulatory T cells (marked by FOXP3 and IL2RA expression) and effector populations may influence disease progression, with regulatory dysfunction potentially permitting unopposed pro-inflammatory and pro-fibrotic T cell responses.

Systemic sclerosis (SSc) is a complex autoimmune disease driven by immune dysregulation, vascular injury, and progressive fibrosis. Although T cells are strongly implicated in disease initiation and progression, defining their functional organization in peripheral blood has been challenging due to technical batch effects, inter-individual variability, and the low abundance of polarized helper subsets in resting PBMCs. In this study, we applied integrative single-cell transcriptomic analysis combined with program-based functional inference to characterize T-cell activation states in scleroderma peripheral blood, providing a robust framework for interpreting immune polarization outside of tissue compartments.

A central finding of this work is that helper T-cell subsets do not segregate into large, discrete clusters in peripheral blood, but instead occupy graded transcriptional continua. This observation is consistent with accumulating evidence that canonical Th1, Th2, Th17, T follicular helper, and regulatory T-cell identities represent dynamic and context-dependent states rather than fixed lineages, particularly in the absence of strong antigenic stimulation.

Despite their limited abundance, each helper T-cell state displayed highly specific and near-orthogonal transcriptional signatures. Th1 cells showed robust expression of IFNG and TBX21, Th17 cells selectively expressed RORC and IL17A, Th2 cells were marked by GATA3 and IL4/IL13, Tfh-like cells expressed BCL6 together with co-stimulatory markers such as PDCD1 and ICOS, and Treg cells expressed FOXP3, IL2RA, and CTLA4. Cytotoxic T cells formed a clearly separable population characterized by high expression of NKG7, GNLY, GZMB, and PRF1, while naive and unpolarized cells displayed minimal effector cytokine expression, serving as an internal negative control. The coherence and specificity of these patterns support the validity of functional-state inference in circulating immune

cells, even in the absence of overt immune activation.

These findings have direct implications for understanding immune contributions to fibrotic disease. Th1-associated interferon signaling, Th17-associated IL-17 signaling, and Tfh-mediated B-cell help have each been implicated in fibrogenic pathways through direct or indirect effects on fibroblasts, endothelial cells, and innate immune populations. The presence of distinct polarization axes in peripheral blood suggests that systemic immune priming may precede or reinforce tissue-specific fibrotic responses. From this perspective, circulating T-cell functional states may reflect an upstream immune milieu that conditions downstream stromal pathology.

Importantly, the analytical framework used here avoids a common pitfall in single-cell studies of autoimmune disease: the assumption that disease-relevant immune biology must manifest as discrete cell clusters. Instead, our results support a model in which pathogenic potential arises from shifts in the balance and intensity of functional programs across a continuum of T-cell states. This view aligns with contemporary immunological frameworks emphasizing plasticity, graded activation, and continuous state transitions rather than rigid lineage boundaries. Such an approach is particularly appropriate for peripheral blood analyses, where effector programs are often partially engaged and cell states are inherently heterogeneous.

The use of scVI-based integration was essential for enabling this analysis at scale. By correcting sample-level batch effects while preserving biological variation, scVI allowed the integration of tens of thousands of cells across hundreds of samples into a unified latent space. This integration was critical for detecting subtle but reproducible polarization signatures that would likely be obscured by technical noise in uncorrected analyses. Coupling this latent representation with conservative quality control and program-based scoring yielded a stable and interpretable view of T-cell functional organization.

Several limitations should be acknowledged. This analysis is restricted to peripheral blood and does not directly capture tissue-resident immune populations, which are likely to exhibit stronger polarization and effector activity in affected organs. In addition, clinical metadata were not incorporated at this stage, precluding direct associations between functional states and disease severity or subtype. Finally, the cross-sectional design limits inference regarding temporal dynamics of immune activation. Nonetheless, the large sample size, rigorous batch correction, and conservative analytical strategy provide a robust foundation for future studies.

In summary, this work demonstrates that functional polarization of T cells in scleroderma peripheral blood can be robustly recovered using integrated single-cell transcriptomics and program-based inference. By revealing coherent Th1, Th17, Th2, Tfh, Treg, and cytotoxic axes without overreliance on discrete clustering, this approach provides a generalizable framework for linking systemic immune states to downstream pathogenic processes. Future studies integrating clinical phenotypes, longitudinal sampling, and tissue-level analyses will be required to determine how these circulating immune programs contribute to fibrosis initiation and progression in systemic sclerosis.

#### Integration with Immune-Fibrosis Pathway

The functional states characterized in this study provide a framework for understanding how systemic immune activation translates to tissue-level pathology in SSc. Recent work has identified immune scarring in B cells following Epstein-Barr virus exposure, characterized by persistent elevation of interferon-stimulated genes and enhanced antigen presentation through upregulated HLA-DR and HLA-DP expression (7,36). These hyperactive B cells would be expected to drive T cell activation through both cognate antigen presentation and costimulatory signaling (36). Our identification of functional T cell states producing pro-fibrotic cytokines (Th2: IL4/IL13; Th17: IL17A) connects this upstream immune dysregulation to the activation of stromal cells. Single-cell profiling of SSc skin fibroblasts has revealed myofibroblast populations with elevated expression of extracellular matrix genes including COL1A1, COL3A1, and ACTA2, which are known transcriptional targets of TGF- $\beta$  and IL-13 signaling pathways (8,9,37). Thus, the immune  $\rightarrow$  stromal axis in SSc likely involves: (1) initial

immune trigger (viral infection, autoantigen exposure), (2) B cell hyperactivation and persistent immune scarring, (3) T cell polarization toward pro-fibrotic states through enhanced antigen presentation, and (4) cytokine-mediated activation of tissue fibroblasts resulting in progressive fibrosis(36,37).

Future studies integrating B-T cell interaction analysis through ligand-receptor inference methods (CellPhoneDB, NicheNet) with patient clinical phenotypes will be valuable for establishing quantitative associations between immune state composition and disease severity. The per-sample functional state composition data we provide (Supplementary Tables S3A-B) can serve as input for such integrative analyses when combined with complete clinical metadata including modified Rodnan skin scores, interstitial lung disease status, and autoantibody profiles. Additionally, spatial transcriptomics approaches applied to SSc tissue biopsies could reveal how circulating T cell states identified here relate to tissue-infiltrating populations and their spatial organization relative to activated fibroblasts and sites of active fibrosis.

### Limitations

Several limitations should be considered when interpreting the present findings. First, this analysis is restricted to peripheral blood and does not directly capture tissue-resident immune populations. T cells within affected organs such as skin and lung are likely to exhibit stronger polarization, higher effector cytokine expression, and more stable lineage commitment than their circulating counterparts. As a result, the functional states identified here should be interpreted as reflecting systemic immune priming rather than local tissue effector responses.

Second, clinical metadata were not incorporated at this stage of analysis. Consequently, direct associations between T-cell functional states and disease severity, clinical subtype, organ involvement, or treatment status could not be evaluated. While the absence of phenotype stratification limits immediate clinical interpretation, it also ensures that the functional-state framework presented here is not confounded by selective subgrouping and can be readily reused when metadata become available.

Third, the cross-sectional nature of the dataset precludes inference about temporal dynamics of immune activation or causal relationships between immune polarization and fibrotic progression. T-cell functional states are inherently dynamic, and longitudinal sampling will be required to determine whether the observed polarization axes represent stable immune imprints or transient activation states.

Finally, although program-based functional inference avoids over-clustering of rare populations, it does not imply mutually exclusive or fixed lineage assignments at the single-cell level. Cells classified as Th1, Th17, Tfh, or regulatory represent dominant transcriptional programs rather than terminally differentiated states. This limitation is intrinsic to peripheral blood analyses and reflects biological plasticity rather than methodological deficiency.

Despite these limitations, the large sample size, rigorous batch correction using scVI, conservative quality control, and coherence of recovered cytokine and transcription factor signatures support the robustness of the presented framework. Together, these considerations position the current study as a foundation for future investigations integrating clinical phenotypes, longitudinal designs, and tissue-level immune profiling.

### Figure Legends

#### Figure 1 | scVI-based integration of PBMCs and extraction of the T-cell compartment

Single-cell transcriptomic integration of peripheral blood mononuclear cells (PBMCs) from systemic sclerosis samples using scVI. (A) Uniform manifold approximation and projection (UMAP) of 67,592 PBMCs embedded in the scVI latent space, demonstrating preservation of major immune lineages. (B) The same UMAP colored by sample identity, showing effective batch mixing across 578 samples following scVI integration. (C) Identification of the T-cell compartment based on marker expression (CD3D, CD3E, TRAC), enabling extraction of a high-confidence T-cell subset ( $n = 14,617$  cells) for

downstream analysis. Together, these panels demonstrate robust batch correction and reliable isolation of T cells from a large, multi-sample PBMC dataset.

#### Figure 2 | T cells organize along graded functional continua rather than discrete helper clusters

Functional organization of T cells in the scVI latent space. (A) UMAP of T cells colored by Leiden clustering (resolution = 0.5), revealing a dominant naive/central memory compartment and a distinct cytotoxic population, while helper subsets do not form large discrete clusters. (B) UMAP colored by inferred functional state based on program-level scoring, highlighting graded polarization states including naive, cytotoxic, Th1, Th2, Th17, T follicular helper (Tfh), regulatory T (Treg), and unpolarized cells. (C) UMAPs colored by representative program scores (naive and cytotoxic), illustrating continuous gradients of activation rather than sharply bounded clusters. These results support a continuum model of helper T-cell polarization in resting peripheral blood.

#### Figure 3 | Cytokine and transcription factor signatures validate T-cell functional states

Cytokine and transcription factor specificity across inferred T-cell functional states. (A) Heatmap showing mean expression of key cytokines and lineage-associated transcription factors for each functional state, demonstrating near-orthogonal signatures for Th1 (IFNG, TBX21), Th17 (RORC, IL17A), Th2 (GATA3, IL4/IL13), Tfh (BCL6), Treg (FOXP3, IL2RA), and cytotoxic T cells (NKG7, GNLY, PRF1). (B) Comparison of IFNG and IL17A expression across functional states, illustrating distinct Th1 and Th17 polarization axes. (C) Representative UMAP gene expression overlays confirming spatial localization of canonical markers within the latent space. These patterns confirm the biological coherence of program-based functional-state inference.

#### Supplementary Figure Legends

##### Supplementary Figure S1 | Additional validation of T-cell marker expression

UMAPs of integrated PBMCs showing expression of additional T-cell marker genes. (A) CD3E expression. (B) TRAC expression. These markers corroborate the identification of the T-cell compartment used for downstream analyses.

##### Supplementary Figure S2 | Functional program score UMAPs for helper and regulatory T-cell states

UMAPs of T cells colored by program-level functional scores for additional helper and regulatory T-cell states. (A) Th1 program score. (B) Th2 program score. (C) Th17 program score. (D) T follicular helper (Tfh) program score. (E) Regulatory T-cell (Treg) program score. These panels illustrate graded polarization patterns across the latent space rather than discrete clustering.

##### Supplementary Figure S3 | T-cell functional program scores across the latent space

UMAPs of T cells colored by program-level functional scores for each inferred functional state. (A) Naive/central memory, (B) Cytotoxic, (C) Th1, (D) Th2, (E) Th17, (F) Tfh, (G) Treg. These panels illustrate graded polarization patterns rather than discrete clustering.

##### Supplementary Figure S4 | Expression of lineage-specific marker genes

Gene-level UMAP overlays showing: (A) IFNG, (B) GZMB, (C) GNLY, (D) FOXP3, (E) CXCR5. These expression patterns validate functional state assignments.

#### Supplementary Table Legends

##### Supplementary Table S1 | Mapping of analytical steps to pipeline scripts

Supplementary Table S1 lists the scripts used at each stage of the computational pipeline, including data ingestion, quality control, latent variable modeling, T-cell extraction, clustering, functional scoring, and result export. This table provides a transparent mapping between the Methods description and the



executable code used to generate all results.

#### Supplementary Table S2A | Mean functional program scores by Leiden cluster (resolution 0.5)

Supplementary Table S2A reports the mean functional program scores for each Leiden cluster identified at resolution 0.5 within the T-cell compartment. Scores correspond to predefined gene programs representing naive/central memory, cytotoxic, Th1, Th2, Th17, Tfh, and Treg functional states, along with the number of cells assigned to each cluster.

#### Supplementary Table S2B | Mean functional program scores by Leiden cluster (resolution 0.8)

Supplementary Table S2B reports the mean functional program scores for each Leiden cluster identified at resolution 0.8 within the T-cell compartment. This higher-resolution clustering illustrates the stability and refinement of functional state assignments across clustering granularities.

#### Supplementary Table S3A | Absolute counts of T-cell functional states per sample

Supplementary Table S3A reports the absolute number of T cells assigned to each functional state for every sample analyzed. Functional states include naive/central memory, cytotoxic, Th1, Th2, Th17, Tfh, Treg, and unpolarized populations.

#### Supplementary Table S3B | Fractional composition of T-cell functional states per sample

Supplementary Table S3B reports the fraction of T cells assigned to each functional state for every sample, normalized by the total number of T cells per sample. These values provide a quantitative summary of cohort-level functional state composition and serve as a reference for future stratification analyses.

### Computational Analysis and Reproducibility

All computational analyses were implemented using a fully reproducible, script-based pipeline orchestrated via a Makefile, with analytical parameters centralized in a YAML configuration file. Briefly, quality-controlled peripheral blood mononuclear cell (PBMC) single-cell RNA sequencing data were subjected to highly variable gene selection prior to latent variable modeling using scVI, enabling batch-aware integration across samples. Latent embeddings and UMAP projections were computed and propagated to the full gene expression space to support downstream analyses. T cells were identified using conservative, marker-based gating and subsequently clustered in the scVI latent space using graph-based methods. Functional gene programs representing naïve, helper, regulatory, and cytotoxic T-cell states were scored in a predefined, hypothesis-driven manner, and cluster-level functional state labels were assigned deterministically based on program enrichment. Sample-level functional state compositions were then quantified and exported for statistical analysis. All analytical steps, corresponding scripts, inputs, outputs, and reproducibility considerations are detailed in Supplementary Table S1, which provides a complete mapping between the Methods and the computational implementation. To ensure methodological reproducibility, analyses were executed on CPU hardware with fixed random seeds and restricted multithreading, and full run provenance, including software versions and configuration checksums, was captured automatically for each execution.

### Data and Code Availability

The single-cell RNA sequencing data analyzed in this study are publicly available from the Gene Expression Omnibus (GEO) under accession number GSE195452. All code used for data processing, analysis, and figure generation is available in a publicly accessible repository and is implemented as a fully reproducible, script-based pipeline orchestrated via a Makefile, with parameters centralized in a YAML configuration file. To ensure methodological reproducibility, analyses were executed using fixed random seeds, CPU-only execution, and restricted multithreading. Full run provenance, including software versions and configuration checksums, is automatically recorded for each execution.

## Author Contributions

GR conceived the study, performed the analysis, interpreted the results, and wrote the manuscript.

## Funding

None.

## Acknowledgments

The author acknowledges the use of AI-based tools during the preparation of this manuscript. OpenAI's ChatGPT (GPT-4) was used to support code prototyping, data analysis scripting, figure generation workflows, and iterative drafting and editing of the main text. Anthropic's Claude (Claude 3.5 Sonnet) was used to generate all supplementary materials, including supplementary figures, supplementary tables, and supplementary notes. All scientific decisions, data interpretations, and conclusions were made by the author, who takes full responsibility for the content of the manuscript.

## Conflict of Interest Statement

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

## References

- Denton CP, Khanna D. Systemic sclerosis. *Lancet*. 2017;390:1685–1699. doi:10.1016/S0140-6736(17)30933-9
- Varga J, Trojanowska M. Fibrosis in systemic sclerosis. *Nat Rev Rheumatol*. 2017;13:325–335. doi:10.1038/nrrheum.2017.37
- Allanore Y, Simms R, Distler JHW, Trojanowska M, Pope J, Denton CP, et al. Systemic sclerosis. *Nat Rev Dis Primers*. 2015;1:15002. doi:10.1038/nrdp.2015.2
- Wynn TA. Cellular and molecular mechanisms of fibrosis. *J Pathol*. 2008;214:199–210. doi:10.1002/path.2277
- Distler JHW, Györfi AH, Ramanujam M, Whitfield ML, Königshoff M, Lafyatis R. Shared and distinct mechanisms of fibrosis. *Nat Rev Rheumatol*. 2019;15:705–730. doi:10.1038/s41584-019-0322-7
- Furie M, Mitoma C, Tsuji G, Kadono T, Uchi H. Role of immune cells in fibrosis. *J Dermatol Sci*. 2017;86:3–9. doi:10.1016/j.jdermsci.2017.01.007
- Wynn TA, Ramalingam TR. Mechanisms of fibrosis: therapeutic translation for fibrotic disease. *Nat Med*. 2012;18:1028–1040. doi:10.1038/nm.2807
- Sakkas LI, Chikanza IC, Platsoucas CD. Mechanisms of T cell activation in systemic sclerosis. *Arthritis Rheum*. 2006;54:2415–2425. doi:10.1002/art.22004
- Fuschiotti P. Role of T cells in systemic sclerosis. *Clin Rev Allergy Immunol*. 2011;40:189–198. doi:10.1007/s12016-010-8213-1
- Parel Y, Aurrand-Lions M, Scheja A, Dayer JM, Roosnek E, Chizzolini C. Th17 and regulatory T cells in systemic sclerosis. *Arthritis Rheum*. 2010;62:1074–1084. doi:10.1002/art.27317
- Radstake TRDJ, van Bon L, Broen J, Wenink M, Santegoets K, Deng Y, et al. Increased Th17 cells in systemic sclerosis are associated with interstitial lung disease. *Arthritis Rheum*. 2009;60:341–352. doi:10.1002/art.24246

Klein S, Kretz CC, Ruland J, Stumpf C, Wollenberg A, Kuhn A. Cytotoxic T cells in systemic sclerosis. *Arthritis Res Ther*. 2015;17:251. doi:10.1186/s13075-015-0776-7

Higashi-Kuwata N, Makino T, Inoue Y, Takeya M, Ihn H. T follicular helper cells in systemic sclerosis. *Arthritis Rheumatol*. 2010;62:364–373. doi:10.1002/art.27145

Antiga E, Quaglini P, Bellandi S, Volpi W, Del Bianco E, Comessatti A, et al. Regulatory T cells in systemic sclerosis. *Autoimmun Rev*. 2010;9:83–87. doi:10.1016/j.autrev.2009.02.020

Mahoney JM, Taroni JN, Martyanov V, Wood TA, Greene CS, Pioli PA, et al. Systems level analysis of systemic sclerosis shows a role for interferon-regulated genes. *Proc Natl Acad Sci USA*. 2008;105:12319–12324. doi:10.1073/pnas.0804914105

Christmann RB, Sampaio-Barros P, Stifano G, Borges CL, de Carvalho CR, Kairalla RA, et al. Association of interferon- and transforming growth factor  $\beta$ -regulated genes with systemic sclerosis-related pulmonary arterial hypertension. *Arthritis Rheum*. 2014;66:2543–2554. doi:10.1002/art.38706

Gourh P, Arnett FC, Tan FK, Assassi S, Divecha D, Paz G, et al. Association of the IL-17 pathway with systemic sclerosis. *Arthritis Rheum*. 2009;60:1953–1962. doi:10.1002/art.24612

Lafyatis R. Transforming growth factor  $\beta$  and fibrosis. *Arthritis Res Ther*. 2014;16:415. doi:10.1186/s13075-014-0415-9

Zhu J, Yamane H, Paul WE. Differentiation of effector CD4 T cell populations. *Annu Rev Immunol*. 2010;28:445–489. doi:10.1146/annurev-immunol-030409-101212

Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol*. 2019;15:e8746. doi:10.15252/msb.20188746

Tung PY, Blischak JD, Hsiao CJ, Knowles DA, Burnett JE, Pritchard JK, et al. Batch effects and the effective design of single-cell gene expression studies. *Nat Methods*. 2017;14:117–124. doi:10.1038/nmeth.4150

Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. *Nat Methods*. 2018;15:1053–1058. doi:10.1038/s41592-018-0229-2

Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*. 2016;352:189–196. doi:10.1126/science.aad0501

Neftel C, Laffy J, Filbin MG, Hara T, Shore ME, Rahme GJ, et al. An integrative model of cellular states, plasticity, and genetics for glioblastoma. *Cell*. 2019;178:835–849. doi:10.1016/j.cell.2019.06.024

Kotliar D, Veres A, Nagy MA, Tabrizi S, Hodis E, Melton DA, et al. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-seq. *Nat Methods*. 2019;16:409–420. doi:10.1038/s41592-019-0393-4

Szabo PA, Miron M, Farber DL. Location, location, location: tissue resident memory T cells in mice and humans. *Nat Immunol*. 2019;20:272–281. doi:10.1038/s41590-019-0319-9

Tabib T, Huang M, Morse C, Papazoglou A, Behera R, Jia M, et al. Cell-type-specific immune dysregulation in idiopathic pulmonary fibrosis and systemic sclerosis-associated interstitial lung disease. *Sci Transl Med*. 2021;13:eabd0318. doi:10.1126/scitranslmed.abd0318

Kim D, Kobayashi T, Voisin B, Jo JH, Sakamoto K, Jin SP, et al. Targeted therapy guided by single-cell transcriptomic analysis in scleroderma. *Nat Commun*. 2019;10:2163. doi:10.1038/s41467-019-09934-1

Gur C, Wang SY, Sheban F, Zada M, Li B, Kharouf F, et al. Lymphocytes promote fibrotic disease progression in systemic sclerosis. *Ann Rheum Dis*. 2020;79:256–266.  
doi:10.1136/annrheumdis-2019-215809

Gur C, et al. LGR5 expressing skin fibroblasts define a major cellular hub perturbed in scleroderma. *Cell*. 2022;185(8):1373-1388.

Shimagami H, et al. Single-cell analysis reveals immune cell abnormalities underlying the clinical heterogeneity of patients with systemic sclerosis. *Nat Commun*. 2025;16:4949.

Tabib T, et al. Myofibroblast transcriptome indicates SFRP2(hi) fibroblast progenitors in systemic sclerosis skin. *Nat Commun*. 2021;12:4384.

Wynn TA. Type 2 cytokines: mechanisms and therapeutic strategies. *Nat Rev Immunol*. 2015;15(5):271-282.

Onishi RM, Gaffen SL. Interleukin-17 and its target genes: mechanisms of interleukin-17 function in disease. *Immunology*. 2010;129(3):311-321.

Distler JHW, et al. Shared and distinct mechanisms of fibrosis. *Nat Rev Rheumatol*. 2019;15(12):705-730.

Ritschel GW. Single-cell signatures of EBV immune scarring in B-cells. *PLoS One*. 2026;submitted. Manuscript Number: PONE-D-26-06471.

Ritschel GW. Single-cell transcriptomic analysis identifies drug-reversible fibroblast activation states in systemic sclerosis. *Front Med (Lausanne)*. 2026;submitted. Manuscript ID: 1801050.