Statistics Innovation

ARTICLE

Open Access

https://doi.org/10.48130/stati-0026-0016

Statistics Innovation **2026**, 3: e015

An interpretable latent disease-axis framework for continuous disease scoring and treatment-response characterization

Yu Wang[1,2,3#], Jiaze Guo[4#], Lei Zuo[1,5#], Shuang Wu[3], Wenqi Pan[3] and Jincan Che[1,3,5*]

[1] *Division of Health Statistics, School of Public Health, Hebei Medical University, Shijiazhuang 050017, China*
[2] *Institute of Statistics and Big Data, Renmin University of China, Beijing 100872, China*
[3] *Beijing Key Laboratory of Topological Statistics and Applications for Complex Systems, Beijing Institute of Mathematical Sciences and Applications, Beijing 101408, China*
[4] *School of Grassland Science, Beijing Forestry University, Beijing 100083, China*
[5] *Hebei Key Laboratory of Environment and Human Health, Shijiazhuang 050017, China*
[#] Authors contributed equally: Yu Wang, Jiaze Guo, Lei Zuo
* Correspondence: chejincan@hebmu.edu.cn (Che J)

Abstract

Many complex diseases exhibit continuous molecular variation and heterogeneous treatment response, which are poorly captured by discrete, category-based models. We propose a latent disease-axis framework for modeling disease progression and treatment effects as continuous processes using high-dimensional molecular data. Using a linear principal component analysis (PCA)-based embedding of transcriptomic profiles, we define a disease axis as a centroid-based geometric direction anchored by healthy controls and untreated disease samples. Projection onto this axis yields a continuous disease score that summarizes the disease severity based on molecular data, while paired pre- and post-treatment samples are represented as displacement vectors that characterize the treatment response in both magnitude and direction relative to the disease axis. To further interpret the biological basis of disease-axis variation, we apply trivariate functional clustering to identify cohort-level gene modules with distinct expression trajectories across health, disease, and post-treatment states. When applied to transcriptomic data relating to systemic lupus erythematosus, this framework reveals a continuous disease structure, heterogeneous treatment-associated shifts, and functionally distinct transcriptional programs with different degrees of reversibility. Overall, the latent disease-axis framework provides an interpretable descriptive strategy for studying continuous disease-related variation and heterogeneous treatment response using high-dimensional molecular data that goes beyond discrete clinical categories.

Keywords: Systemic lupus erythematosus, Transcriptomics, Functional clustering, Latent disease axis, Treatment-response heterogeneity

Citation: Wang Y, Guo J, Zuo L, Wu S, Pan W, et al. 2026. An interpretable latent disease-axis framework for continuous disease scoring and treatment-response characterization. *Statistics Innovation* 3: e015 https://doi.org/10.48130/stati-0026-0016

Introduction

Many complex diseases are characterized by high-dimensional molecular measurements, gradual progression, and heterogeneous responses to therapy[1–4]. Yet, they are often analyzed using statistical models that treat disease states and treatment outcomes as discrete categories, such as case–control contrasts or responder versus non-responder classifications[5–7]. Although such formulations are convenient, they can obscure continuous variations in disease severity, fail to capture intermediate or transitional states, and provide limited resolution for quantifying individual-level treatment effects[8]. In particular, when molecular measurements are used to trace a continuum of latent disease states rather than sharply separated groups, methods that rely solely on categorical labels may underutilize the geometric structure of the data and offer an incomplete description of disease dynamics[9,10].

High-throughput transcriptomic profiling has enabled detailed characterization of molecular heterogeneity in many immune-mediated diseases, including systemic lupus erythematosus (SLE)[11,12]. Prior studies have identified disease-associated gene signatures, molecular subtypes, and pathway-level alterations using differential expression analysis, clustering, or supervised classification[13,14]. However, these approaches typically impose discrete boundaries on samples and are not designed to model disease severity as a continuous quantity or to represent treatment response as a gradual, patient-specific process[15]. In diseases such as SLE, in which molecular states often span a spectrum between health and active disease and therapeutic effects vary markedly across individuals[16], there is a need for statistical frameworks that can capture continuous disease-related variation and heterogeneous treatment dynamics within a unified representation[17].

In this work, we propose a latent disease-axis framework to describe disease progression and treatment response as continuous geometric processes in a low-dimensional transcriptomic space. Using a linear principal component analysis (PCA)-based embedding of high-dimensional transcriptomic profiles, we define a disease axis as a centroid-based direction anchored by healthy controls and untreated disease samples. Projection onto this axis yields a continuous disease score that summarizes the disease severity based on molecular data, while paired pre- and post-treatment samples are represented as displacement vectors that capture individualized treatment trajectories in both magnitude and direction. To further interpret the biological basis of disease-axis variation, we apply trivariate functional clustering (3FunClu) to identify cohort-level gene modules with distinct expression trajectories across health, disease, and post-treatment states. When applied to transcriptomic data relating to SLE, this framework reveals a continuous disease structure, heterogeneous treatment-associated shifts, and functionally distinct transcriptional programs with different degrees of reversibility. Overall, the latent disease-axis framework provides an interpretable descriptive strategy for studying continuous disease-related variation using high-dimensional molecular data that goes beyond discrete clinical categories.

https://doi.org/10.48130/stati-0026-0016

Materials and methods

Data source and preprocessing

RNA-seq expression profiles were obtained from the Gene Expression Omnibus (GEO; accession GSE295130), which includes data pertaining to peripheral blood mononuclear cell (PBMC) samples from healthy controls (HC), untreated SLE patients, and matched post-treatment samples (SLE_post). The analyzed dataset contained 19 PBMC samples, including 7 HC, 6 untreated SLE, and 6 SLE_post samples. The paired treatment–displacement analysis used six matched pre/post-treatment pairs: D1–B1, D2–B2, D3–B3, D5–B5, D10–B10, and D11–B11. Let $X \in \mathbb{R}^{n\times p}$ denote the processed FPKM expression matrix, where FPKM stands for Fragments Per Kilobase of exon model per Million mapped fragments, n is the number of samples and p is the number of genes. In the processed matrix, p = 66,332 genes before filtering.

Expression values were transformed using a log-scale normalization:

$$X^*_{ij} = \log_2(X_{ij}+1),$$

to stabilize the variance and reduce the influence of extreme counts[18,19]. Genes were first filtered using a global low-expression criterion, retaining those with FPKM ≥ 1 in at least four samples. Consequently, 17,514 genes were retained from the initial 66,332 genes. Genes with non-positive or near-zero variance after log transformation were also excluded before downstream scaling[20].

For the primary PCA-based disease-axis analysis, genes passing the low-expression filter were ranked by their across-sample variance on the log-transformed scale, and the top 2,000 highly variable genes (HVGs) were used. The same workflow was repeated in the sensitivity analyses using 1,000 and 3,000 HVGs.

The selected HVG matrix was standardized across samples:

$$Z_{ij} = \frac{X^*_{ij}-\mu_j}{\sigma_j},$$

where μ_j and σ_j denote the sample mean and standard deviation of gene j, respectively. Both μ_j and σ_j were computed across all analyzed samples after log transformation and HVG selection.

Each sample was thus represented by a high-dimensional standardized expression vector $z_i \in \mathbb{R}^p$. Sample labels (HC, SLE, and SLE_post) were derived from metadata annotations and were used only to define the HC and untreated-SLE reference centroids, identify paired pre/post-treatment samples, and annotate visualizations. No supervised classifier, training-test split, or cross-validation procedure was fitted in this workflow. Because sex, age, and batch covariates were not reliably available in the processed matrix used for reanalysis, no formal covariate adjustment was performed; residual demographic or technical confounding was therefore treated as a limitation rather than excluded by design.

Latent disease-axis modeling

Latent embedding via PCA

To obtain a low-dimensional representation of transcriptomic states, PCA was applied to the standardized HVG expression matrix Z, treating samples as observations and genes as variables. PCA provides an orthogonal linear transformation that captures dominant sources of variance in the data while reducing dimensionality. In the high-dimensional, low-sample-size setting of our study, PCA was used as an unsupervised dimensionality reduction technique rather than as a predictive model. For PCA-based disease-axis analysis, we used the top 2,000 HVGs ranked by across-sample variance on the $\log_2$(FPKM+1) scale. The selected HVG matrix was z-scored across samples (centered and scaled per gene) prior to PCA.

Let $y_i = W^T z_i \in \mathbb{R}^k$ denote the latent embedding of sample i, where W denotes the loading vectors of the leading principal components. In this study, we used the first two principal components as the primary visualization space ($k = 2$), which balances interpretability and direct geometric visualization of disease-state separation and paired treatment-associated displacement[21]. In the primary top-2,000-HVG PCA, PC1 explained 27.40% of the variance and PC2 explained 16.98%, yielding a cumulative explained variance of 44.38%.

The primary disease axis, HC-centered disease scores, and paired treatment-response vectors described below were computed in the PC1–PC2 plane for visualization and interpretation. To assess whether the conclusions depended on the two-dimensional representation, the same centroid-based axis construction and HC-centered scoring procedure were repeated in higher-dimensional PCA subspaces, including fixed $k \in \{2,3,5\}$ and cumulative-variance-based dimensions.

Centroid-defined latent disease axis and continuous disease score

The latent disease axis is defined as a centroid-based geometric direction from the healthy reference state toward untreated SLE within the PCA latent space. Let I_{HC} and I_{SLE} denote the index sets of HC and untreated SLE samples, respectively. The corresponding group centroids in latent space are

$$\mu_{\mathrm{HC}} = \frac{1}{|I_{\mathrm{HC}}|}\sum_{i\in I_{\mathrm{HC}}} y_i, \quad \mu_{\mathrm{SLE}} = \frac{1}{|I_{\mathrm{SLE}}|}\sum_{i\in I_{\mathrm{SLE}}} y_i.$$

The disease axis is defined as the unit-length vector

$$v = \frac{\mu_{\mathrm{SLE}}-\mu_{\mathrm{HC}}}{\|\mu_{\mathrm{SLE}}-\mu_{\mathrm{HC}}\|},$$

which points from the healthy centroid toward the untreated SLE centroid. SLE_post samples were not used in constructing this axis, thereby avoiding information leakage.

Each sample is assigned a continuous disease score by projection onto the disease axis. To make the score explicitly relative to the healthy reference, the projection was centered at the HC centroid:

$$s_i = (y_i - \mu_{\mathrm{HC}})^T v.$$

Larger values of s_i indicate transcriptional states more similar to the untreated SLE centroid. Because the score is centered at μ_{HC}, HC samples are interpreted relative to the healthy reference centroid, and movement toward smaller s_i values corresponds to a shift toward the healthy direction along the disease axis. For visualization and groupwise comparison, disease scores were standardized across all samples:

$$\tilde{s}_i = \frac{s_i - \bar{s}}{\mathrm{SD}(s)}$$

This standardization was used only for visualization and groupwise display; treatment-response direction and paired displacement were interpreted using the raw HC-centered score s_i.

Vector-based modeling of treatment response

For each patient p with paired pre- and post-treatment samples, let $y_{p,\mathrm{SLE}}$ and $y_{p,\mathrm{SLE\_post}}$ denote the PCA latent coordinates of the untreated and post-treatment states, respectively. Treatment response is modeled as a displacement vector:

$$\boldsymbol{\delta}_p = \boldsymbol{y}_{p,\mathrm{SLE\_post}} - \boldsymbol{y}_{p,\mathrm{SLE}}.$$

The scalar change in disease score along the axis is defined as the difference between the post-treatment and untreated disease scores. Under the revised HC-centered disease-score definition, this quantity is computed as

$$\Delta s_p = s_{p,\mathrm{SLE\_post}} - s_{p,\mathrm{SLE}} = \langle \boldsymbol{y}_{p,\mathrm{SLE\_post}} - \boldsymbol{\mu}_{\mathrm{HC}}, \boldsymbol{v} \rangle - \langle \boldsymbol{y}_{p,\mathrm{SLE}} - \boldsymbol{\mu}_{\mathrm{HC}}, \boldsymbol{v} \rangle = \langle \boldsymbol{\delta}_p, \boldsymbol{v} \rangle.$$

Thus, although the disease score itself is centered at the HC centroid, the paired score change remains equivalent to the projection of the treatment displacement vector onto the disease axis. Under this convention, $\Delta s_p < 0$ indicates movement toward the healthy reference direction along the disease axis.

To explicitly characterize the alignment between the treatment vector and the disease axis, we compute the cosine similarity:

$$\cos(\theta_p) = \frac{\langle \delta_p, v \rangle}{\|\delta_p\|_2 \|v\|_2} = \frac{\langle \delta_p, v \rangle}{\|\delta_p\|_2},$$

Because v is a unit vector pointing from the HC centroid toward the untreated SLE centroid, negative cosine similarity indicates that the paired treatment displacement is oriented opposite to the disease axis, i.e., toward the healthy reference direction.

Functional gene modules and enrichment analysis

3FunClu of gene-expression trajectories

To characterize coordinated transcriptional programs across disease and treatment states, we applied 3FunClu[22–24] to classify genes according to the similarity of their expression trajectories across HC, SLE, and SLE_post. This analysis treated each gene as a functional unit whose expression pattern was jointly modeled over the three disease-related conditions, thereby defining gene modules according to trajectory similarity rather than pairwise co-expression correlation.

For 3FunClu, we used all 17,514 genes that passed the low-expression filter (retaining genes with FPKM ≥ 1 in at least four samples), rather than restricting our analysis to the top 2,000 HVGs used in the PCA-based disease-axis analysis. This is because 3FunClu aims to capture coordinated transcriptional programs across the full transcriptome, including moderately expressed genes that may be functionally important but would be excluded by HVG selection. Indeed, only 55 of the 61 identified modules contained HVGs, supporting the use of the full filtered gene set for functional clustering. In contrast to PCA, non-negative log-transformed expression values were used directly as input to the finite-mixture model, without additional z-scoring, and expression trajectories were modeled using condition-specific mean expression values across HC, SLE, and SLE_post rather than individual sample profiles.

For each gene g, let

$$\boldsymbol{x}_g = \left(\boldsymbol{x}_g^{\mathrm{HC}}, \boldsymbol{x}_g^{\mathrm{SLE}}, \boldsymbol{x}_g^{\mathrm{SLE\_post}}\right)$$

denote its concatenated expression profile across the three conditions. Because the FunClu mean model is based on power-law curve fitting, the clustering analysis was performed on non-negative log-transformed expression values after gene filtering. The expression trajectory of each gene was modeled using a finite-mixture likelihood:

$$\mathcal{L}(\Theta) = \sum_{g=1}^{m} \log\left[\sum_{\ell=1}^{K} \pi_\ell \mathcal{N}\left(\boldsymbol{x}_g \mid \boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell\right)\right],$$

where π_ℓ is the prior probability of module ℓ, $\boldsymbol{\mu}_\ell$ denotes the module-specific mean trajectory across HC, SLE, and SLE_post, and $\boldsymbol{\Sigma}_\ell$ denotes the residual covariance structure.

For module ℓ in condition c, the mean expression curve was approximated by the power function

$$\mu_\ell^c(h) = \alpha_{\ell c} h^{\beta_{\ell c}}, \qquad c \in \{\mathrm{HC}, \mathrm{SLE}, \mathrm{SLE\_post}\},$$

where $\alpha_{\ell c}$ and $\beta_{\ell c}$ describe the magnitude and state-index-dependent shape of the trajectory, respectively. The optimal number of functional modules was selected using the Bayesian information criterion (BIC). After model fitting, each gene was assigned to the module with the highest posterior probability. Modules showing marked divergence between HC and SLE or partial reversal from SLE to SLE_post were prioritized for downstream functional enrichment analysis.

Gene Ontology (GO) enrichment of FunClu-derived modules

To interpret the biological functions of FunClu-derived gene modules, GO enrichment analysis was performed for the biological process (BP), molecular function (MF), and cellular component (CC) categories.

Over-representation analysis was conducted using a hypergeometric framework, and p values were corrected using the Benjamini–Hochberg false discovery rate (FDR) procedure.

GO terms with adjusted $p < 0.05$ were considered significantly enriched.

Modules showing clear HC–SLE divergence or partial SLE-to-SLE_post reversal in their fitted trajectories were prioritized for interpretation to connect trajectory-defined gene modules with immune and inflammatory categories relevant to SLE.

Results

A latent disease axis captures continuous disease severity and treatment-associated shifts

We first examined whether transcriptomic variation across HC, untreated SLE patients, and SLE_post could be represented within a continuous low-dimensional structure. PCA of the top 2,000 HVGs revealed that the first two components captured 27.40% and 16.98% of the total variance, respectively (cumulative 44.38%; Fig. 1). In this embedding, we observed a broad separation between HC and disease-associated samples, while also noting substantial overlap among groups.

This overlap indicates that molecular disease states do not form strictly discrete clusters, but instead occupy a continuum within the latent space.

Using group centroids defined by HC and untreated SLE samples, we constructed a latent disease axis and projected all samples onto this direction. Each sample was assigned a HC-centered continuous disease score. The resulting scores exhibited a clear ordering from the healthy reference toward more SLE-like transcriptional states (Fig. 2a). Untreated SLE samples showed significantly higher scores than HC, whereas SLE_post samples were distributed at intermediate positions, reflecting partial shifts toward the healthy end of the axis.

Importantly, disease-axis scores overlapped across clinical categories, underscoring that molecular disease severity varies continuously rather than discretely. This continuous representation provides a compact descriptive summary of the 19 PBMC transcriptomes while preserving inter-individual heterogeneity, which is obscured by binary case–control contrasts.

Paired analysis of the six matched pre- and post-treatment pairs demonstrated that treatment was associated with a reduction in HC-centered disease-axis scores ($\Delta s_p < 0$) in five out of six patients, although the magnitude of change varied markedly across

https://doi.org/10.48130/stati-0026-0016

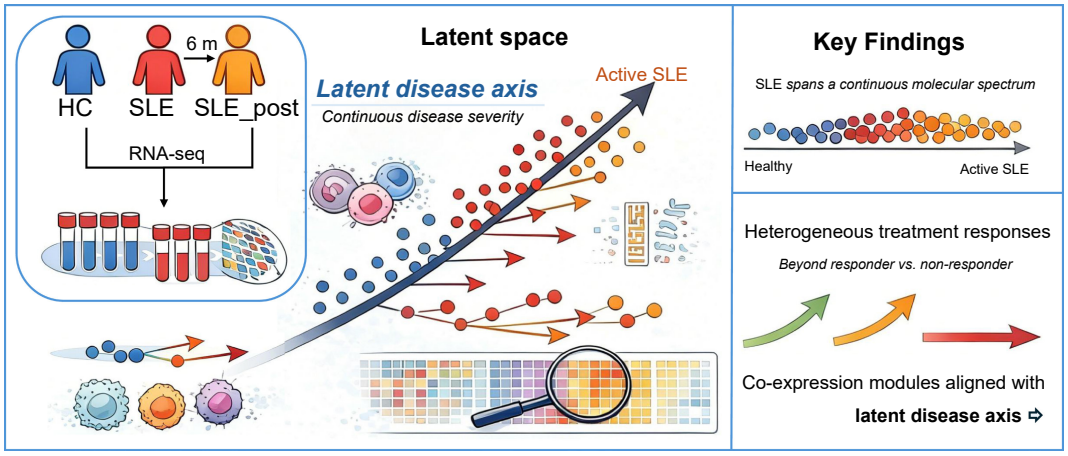

Fig. 1 Conceptual framework of a latent disease axis for SLE.

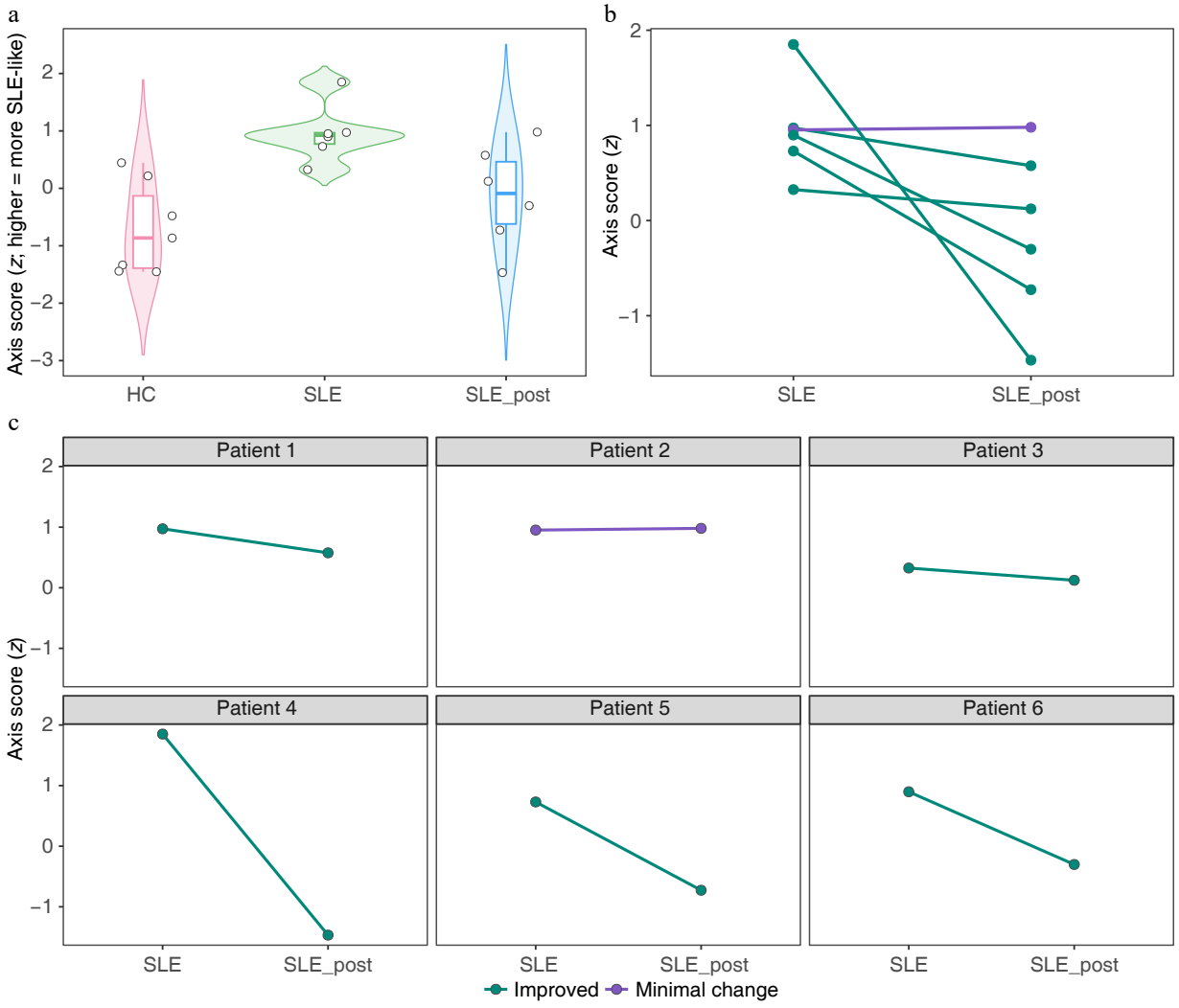

Fig. 2 Disease-axis projection and paired treatment shifts. (a) Distribution of disease-axis projection scores for healthy controls (HC), untreated SLE patients (SLE), and post-treatment SLE samples (SLE_post). Higher scores indicate a more SLE-like transcriptional state. (b) Paired shifts along the disease axis for individual patients, comparing pre-treatment (SLE) and post-treatment (SLE_post) samples. Each line represents one patient. (c) Individual-level treatment trajectories, highlighting heterogeneous responses ranging from marked improvement to minimal change.

individuals (Fig. 2b). Together, these findings support the latent disease axis as an interpretable one-dimensional summary of disease severity that is sensitive to treatment-associated transcriptional shifts without imposing categorical boundaries.

Patient-specific treatment trajectories reveal heterogeneous, direction-aware responses

Although disease-axis scores provide a scalar summary of molecular severity, they do not capture the full structure of

treatment-induced changes. To characterize individualized responses, we modeled treatment effects as displacement vectors between paired pre- and post-treatment samples in the latent space.

Visualization of patient-specific treatment trajectories revealed pronounced heterogeneity across the six patients (Figs 2c and 3b). Although five individuals exhibited displacement vectors oriented toward the healthy region of the latent space, consistent with reductions in HC-centered disease-axis scores, one patient showed a minor shift away from the healthy centroid despite comparable baseline severity.

Notably, several treatment trajectories contained substantial components orthogonal to the disease axis, indicating transcriptional changes that are not fully captured by a single severity dimension. These orthogonal components suggest that treatment may induce molecular responses related to compensatory pathways, off-target effects, or individual-specific regulatory programs rather than uniform movement along a disease progression axis.

By explicitly modeling both the magnitude and direction of transcriptional change, the vector-based framework reveals dimensions of treatment response that are obscured by scalar pre–post comparisons. This direction-aware representation highlights that treatment response in SLE is inherently heterogeneous and patient-specific.

Latent-space geometry links disease progression and treatment effects

To contextualize HC-centered disease-axis scores within the broader structure of transcriptomic variation, we visualized samples and treatment trajectories directly in the latent space defined by the leading principal components (Fig. 3a). In this space, untreated SLE samples occupied regions distinct from HCs, whereas SLE_post samples exhibited heterogeneous displacements.

Treatment trajectories oriented toward the HC region tended to correspond to reductions in disease-axis scores, whereas trajectories oriented away from this region or orthogonal to the disease axis were associated with minimal improvement. This geometric relationship clarifies how scalar disease scores arise from underlying multivariate expression patterns.

The latent-space visualization highlights the complementary roles of the disease axis and the full embedding. While the disease axis captures a dominant direction of variation associated with disease severity, the full latent space preserves additional orthogonal dimensions that contribute to individual-specific treatment responses. Thus, the geometric framework integrates unidimensional severity scoring with multivariate representation of heterogeneous molecular dynamics.

Functional trajectory clustering and pathway enrichment provide a functional interpretation of disease-axis variation

To characterize the cohort-level transcriptional programs associated with disease and treatment, we performed functional clustering analysis (3FunClu), in which genes with similar modeled expression trajectories were grouped into the same module. This analysis identified four major module classes across HC, untreated SLE, and SLE_post (Fig. 4), defined based on their expression trajectory patterns as follows. Partial recovery modules (PRMs) were characterized by downregulated expression in untreated SLE compared to HC, followed by partial increase after treatment that did not return to HC levels. Full recovery modules (FRMs) showed altered expression in untreated SLE but returned to levels comparable to HC after treatment, indicating complete reversibility. Persistent modules (PMs) exhibited altered expression in untreated SLE with no significant restoration after treatment, remaining in a disease-like state. Stable modules (SMs) displayed unchanged expression across all three conditions (HC, SLE, and SLE_post), showing no disease- or treatment-associated variation. PRMs showed marked downregulation in untreated SLE and partial restoration after treatment, indicating incomplete recovery of disease-suppressed programs. FRMs displayed clear disease-associated perturbation followed by substantial recovery toward

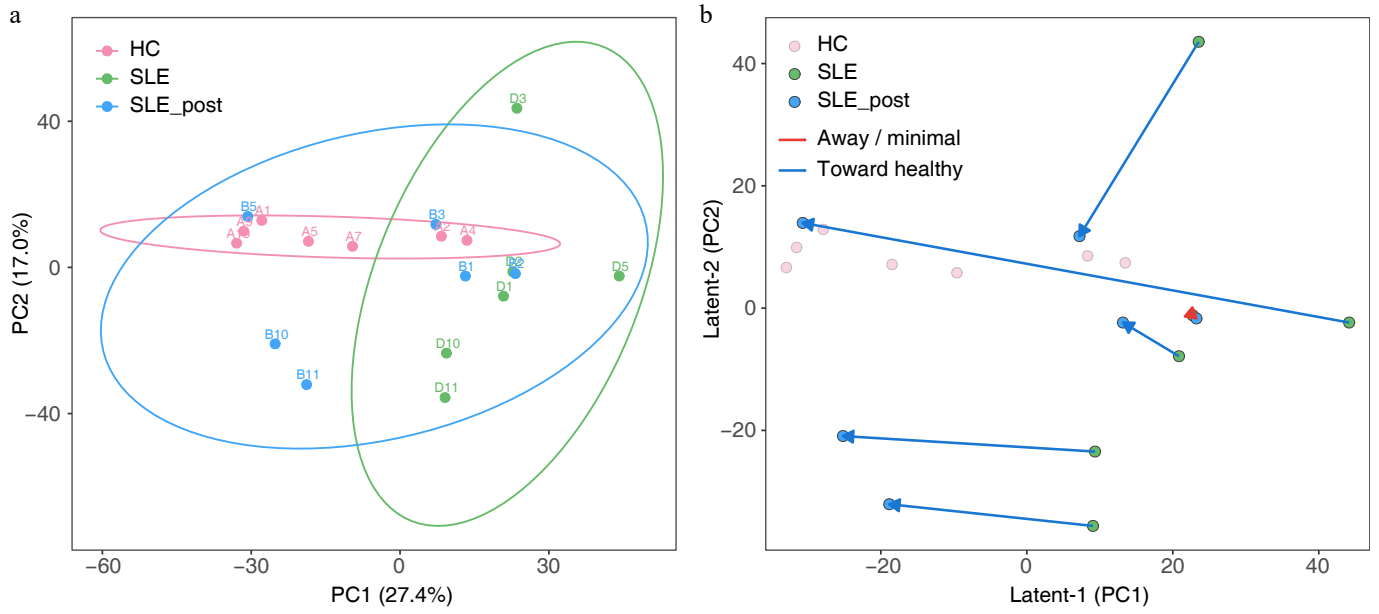

Fig. 3 Representation analysis revealing treatment-associated trajectories in the latent space. (a) PCA of normalized gene expression across HC, SLE (pre-treatment), and SLE_post (post-treatment). Ellipses denote 95% confidence regions. (b) Latent manifold embedding showing individual treatment vectors $\delta_i = z_{i,\mathrm{SLET}} - z_{i,\mathrm{SLE}}$. Vectors are colored by their projection onto the disease axis v (blue: $\Delta s_i > 0$, red: $\Delta s_i \leq 0$).

https://doi.org/10.48130/stati-0026-0016

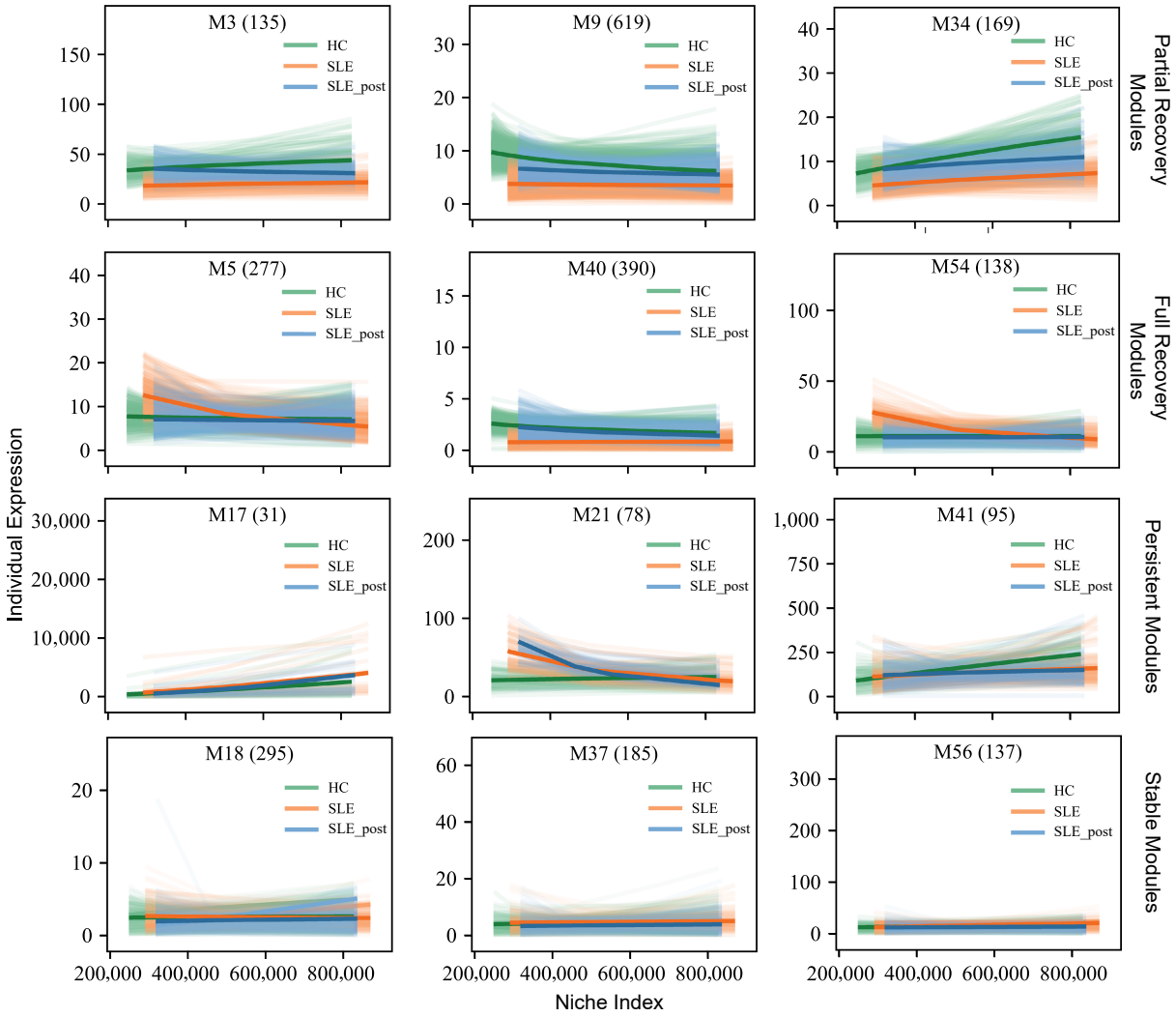

Fig. 4 Representative trajectory-defined modules identified by trivariate functional clustering across healthy controls (HC), untreated SLE, and SLE_post (post-treatment SLE).

HC-like levels, suggesting that these represent the most reversible components of the shift. In contrast, PMs remained markedly dysregulated after treatment, explaining why SLE_post samples often remain intermediate along the disease axis.

GO enrichment analysis further showed that these classes correspond to biologically distinct layers of the disease continuum (Fig. 5). PRMs were mainly enriched in MAPK regulation, mitochondrial gene expression, and innate immune response, indicating that metabolic support and stress adaptation are only partially restored. By contrast, FRMs were enriched in transcriptional regulation, intracellular transport, and adaptive immune response, revealing that the most reversible component is concentrated in coordinated regulatory systems.

Conversely, PMs were predominantly enriched in oxidative phosphorylation, mitochondrial ATP synthesis, and chromatin-level processes, indicating a structured persistent state. Lastly, SMs were associated with RNA processing and structural components, staying relatively stable across groups.

Together, the module-level and pathway-level analyses provide functional grounding for the latent disease-axis framework, linking abstract geometric representations of disease progression and treatment response to interpretable molecular programs. This functional layering explains the structured nature of treatment response observed in the latent transcriptomic space.

Discussion

In this study, we proposed a latent disease-axis framework to describe disease severity and treatment response as continuous geometric processes in a transcriptomic space[8]. By projecting high-dimensional gene expression profiles into a low-dimensional unsupervised space and defining a biologically anchored disease axis, heterogeneous molecular states are mapped into a unified coordinate system that supports both sample-level disease scoring and patient-level treatment-trajectory analysis. When applied to SLE, this framework reveals a continuous spectrum of disease-associated transcriptional variation and heterogeneous, direction-aware treatment trajectories that are not readily captured by discrete modeling approaches[1,11,16,25].

A central insight of the proposed framework is that molecular disease states are more appropriately represented as points along a continuum rather than as members of sharply separated categories[5,26]. Although untreated SLE samples exhibited higher

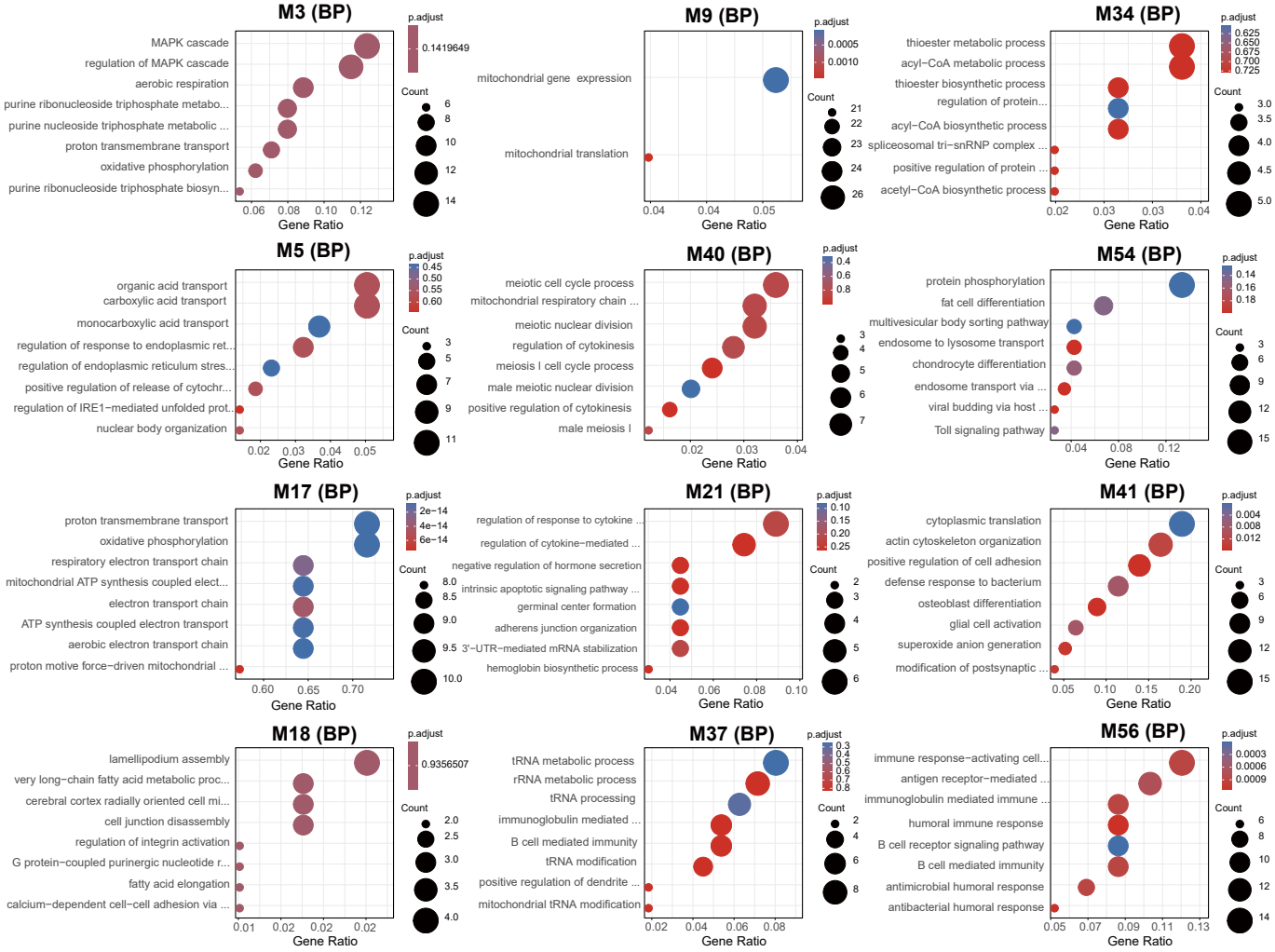

Fig. 5 GO biological process enrichment of PRM, FRM, PM, and SM reveals functional layers of the disease continuum.

HC-centered disease-axis scores on average than HC, the substantial overlap across groups indicates that binary case–control contrasts obscure intermediate and transitional states. By explicitly modeling disease severity as a continuous coordinate, the latent disease axis preserves inter-individual variability while providing an interpretable descriptive characterization of the transcriptomic landscape[27].

From a methodological standpoint, the choice of a linear PCA-based embedding serves as a deliberately simple and auditable framework for high-dimensional, low-sample-size (HDLSS) datasets. While sophisticated non-linear manifold learning and trajectory inference (TI) methods can capture complex geometries, a linear axis anchored by group centroids provides direct interpretability for small paired cohorts. Our sensitivity analyses—including HVG count sweeps, leave-one-out (LOO) stability checks, and higher-dimensional PCA extensions—demonstrated that the group ordering (HC < SLE_post < SLE) and the directional conclusions for paired treatment shifts (Δs_p) remain qualitatively stable under varying analysis choices.

Importantly, the disease axis is not equivalent to any single principal component[28]. Although PCA provides an unsupervised embedding that captures dominant sources of variance, the disease axis is defined as a biologically anchored direction based on group centroids[8]. This construction separates the global variance structure from disease-relevant progression and allows quantification of disease severity independently of the axes of maximal variance. As a result, the framework avoids overinterpretation of unsupervised components while retaining the interpretability and stability of linear dimensionality reduction techniques[29].

Beyond scalar severity scores, the framework explicitly models treatment response as a vector-valued process in a latent space. This geometric formulation reveals that treatment effects in SLE are heterogeneous not only in magnitude but also in direction[30]. While five out of six patients exhibited transcriptional shifts toward the healthy end of the disease axis, one individual showed limited movement or changes oriented orthogonally to the axis. Such orthogonal components may reflect compensatory responses, off-target effects, or patient-specific regulatory programs that are not captured by a single dimension of disease severity. By retaining these directional features, the vector-based representation provides a richer characterization of treatment-associated molecular dynamics than that achieved using scalar pre–post comparisons alone[31].

The integration of the geometric framework with 3FunClu provided deep biological insights into the components of the disease axis. This cohort-level functional layer was not intended to redefine the disease axis, but to interpret the biological programs underlying the disease continuum identified at the sample level. By classifying genes into PRM, FRM, PM, and SM modules based on their HC–SLE–SLE_post trajectories, we found that the disease continuum is composed of distinct functional layers with varying

https://doi.org/10.48130/stati-0026-0016

reversibility. The finding that FRMs are broadly aligned with treatment-responsive shifts, whereas PMs remain dysregulated after treatment, helps explain why patients often occupy an intermediate position along the disease axis rather than return to a healthy baseline. By utilizing a linear embedding and deterministic geometric projections, the approach avoids overparameterization and minimizes susceptibility to overfitting.

Several limitations of our approach should be acknowledged. First, this study serves as a single-cohort proof-of-concept descriptive analysis. Due to the absence of an independent external SLE cohort, the generalizability of the specific axis direction and gene weights requires future validation. Second, the disease axis is defined using cohort-specific centroids and may depend on sample composition. Alternative anchoring strategies, such as external references or longitudinal baselines, could be explored. Third, the current analysis does not formally adjust for demographic or batch covariates as they were not available in the processed matrix; while the linear choice enhances interpretability, complementary nonlinear methods may capture additional structures at the cost of reduced transparency.

In summary, this work reframes disease progression and treatment response as continuous geometric processes in the latent transcriptomic space. By integrating unsupervised embedding, geometric projection, vector-based treatment modeling, and cohort-level functional trajectory clustering, the latent disease-axis framework provides an interpretable statistical approach for analyzing heterogeneous disease states and individualized treatment effects. This paradigm may facilitate a more nuanced modeling of disease dynamics beyond discrete clinical categories and has potential applicability across a broad range of complex diseases.

Conclusions

We proposed a latent disease-axis framework that describes disease severity and treatment response as continuous and direction-aware processes in transcriptomic space. Using a simple PCA-based embedding and a centroid-defined disease axis, the framework provides an interpretable way to summarize molecular disease position at the sample level and to characterize patient-specific treatment-associated shifts through paired displacement vectors. When applied to SLE, the analysis revealed a continuous disease structure rather than sharply separated molecular states, together with heterogeneous treatment trajectories across patients.

By further integrating 3FunClu and GO enrichment analysis, we showed that the disease continuum is functionally composed of transcriptional programs with different degrees of reversibility, including partially reversible, fully reversible, persistent, and relatively stable components. Persistent modules enriched in mitochondrial, energetic, and chromatin-related processes suggest that SLE_post samples often represent an intermediate molecular state rather than a complete return to health.

Overall, this framework provides a transparent and descriptive strategy for linking sample-level geometric variation with cohort-level functional interpretation in high-dimensional molecular data. Although demonstrated here in a small single-cohort SLE dataset, this strategy may be useful for studying continuous disease-related variation and heterogeneous treatment response in other molecular settings.

Author contributions

The authors confirm their contributions to this study as follows: study conception and design, data collection, analysis and interpretation of results, manuscript preparation: Che J; analysis and interpretation of results, additional analyses, result interpretation, manuscript revision: Wang Y, Guo J, Zuo L; manuscript preparation: Wu S, Pan W. All authors reviewed the results and approved the final version of the manuscript.

Data availability

The datasets analyzed during the current study are available in the NCBI Gene Expression Omnibus (GEO) repository under accession number GSE295130 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE295130).

Acknowledgments

This work was funded by the Postdoctoral Foundation of Hebei Medical University (NO.30705010103). The authors are thankful to the Editorial Office for their help in the submission and review process, and the two reviewers for their thoughtful and constructive comments.

Conflict of interest

The authors declare that they have no conflict of interest.

Dates

Received 26 December 2025; Revised 9 June 2026; Accepted 20 July 2026; Published online 4 September 2026

References

[1] Brodin P, Davis MM. 2017. Human immune system variation. *Nature Reviews Immunology* 17(1):21–29

[2] Hasin Y, Seldin M, Lusis A. 2017. Multi-omics approaches to disease. *Genome Biology* 18(1):83

[3] Wu S, Liu X, Dong A, Gragnoli C, Griffin C, et al. 2023. The metabolomic physics of complex diseases. *Proceedings of the National Academy of Sciences of the United States of America* 120(42):e2308496120

[4] Che J, Wang Y, Feng L, Gragnoli C, Griffin C, et al. 2025. High-order interaction modeling of tumor-microenvironment crosstalk for tumor growth. *Physics of Life Reviews* 54:11–23

[5] Haghverdi L, Buettner F, Theis FJ. 2015. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* 31(18):2989–2998

[6] Hobson P, Lovell BC, Percannella G, Vento M, Wiliem A. 2015. Benchmarking human epithelial type 2 interphase cells classification methods on a very large dataset. *Artificial Intelligence in Medicine* 65(3):239–250

[7] Villani AC, Satija R, Reynolds G, Sarkizova S, Shekhar K, et al. 2017. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* 356(6335):eaah4573

[8] Moon KR, van Dijk D, Wang Z, Gigante S, Burkhardt DB, et al. 2019. Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology* 37(12):1482–1492

[9] Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, et al. 2014. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology* 32(4):381–386

[10] Enciso J, Pelayo R, Villarreal C. 2019. From discrete to continuous modeling of lymphocyte development and plasticity in chronic diseases. *Frontiers in Immunology* 10:1927

[11] Nagafuchi Y, Shoda H, Fujio K. 2019. Immune profiling and precision medicine in systemic lupus erythematosus. *Cells* 8(2):140

[12] Nakano M, Iwasaki Y, Fujio K. 2021. Transcriptomic studies of systemic lupus erythematosus. *Inflammation and Regeneration* 41(1):11

[13] Junta CM, Sandrin-Garcia P, Fachin-Saltoratto AL, Mello SS, Oliveira RD, et al. 2009. Differential gene expression of peripheral blood mononuclear cells from rheumatoid arthritis patients may discriminate immunogenetic, pathogenic and treatment features. *Immunology* 127(3):365–372

[14] Nehar-Belaid D, Hong S, Marches R, Chen G, Bolisetty M, et al. 2020. Mapping systemic lupus erythematosus heterogeneity at the single-cell level. *Nature Immunology* 21(9):1094–1106

[15] Toro-Domínguez D, Martorell-Marugán J, Goldman D, Petri M, Carmona-Sáez P, et al. 2018. Stratification of systemic lupus erythematosus patients into three groups of disease activity progression according to longitudinal gene expression. *Arthritis & Rheumatology* 70(12):2025–2035

[16] Barturen G, Babaei S, Català-Moll F, Martínez-Bueno M, Makowska Z, et al. 2021. Integrative analysis reveals a molecular stratification of systemic autoimmune diseases. *Arthritis & Rheumatology* 73(6):1073–1085

[17] Reicher L, Shilo S, Godneva A, Lutsker G, Zahavi L, et al. 2025. Deep phenotyping of health–disease continuum in the Human Phenotype Project. *Nature Medicine* 31(9):3191–3203

[18] Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* 15(12):550

[19] Law CW, Chen Y, Shi W, Smyth GK. 2014. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* 15(2):R29

[20] Bourgon R, Gentleman R, Huber W. 2010. Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences of the United States of America* 107(21):9546–9551

[21] Ringnér M. 2008. What is principal component analysis? *Nature Biotechnology* 26(3):303–304

[22] Kim BR, Zhang L, Berg A, Fan J, Wu R. 2008. A computational approach to the functional clustering of periodic gene-expression profiles. *Genetics* 180(2):821–834

[23] Kim BR, McMurry T, Zhao W, Wu R, Berg A. 2010. Wavelet-based functional clustering for patterns of high-dimensional dynamic gene expression. *Journal of Computational Biology* 17(8):1067–1080

[24] Kim BR, Littell RC, Wu R. 2006. Clustering periodic patterns of gene expression based on Fourier approximations. *Current Genomics* 7(3):197–203

[25] Lincoln MR, Connally N, Axisa PP, Gasperi C, Mitrovic M, et al. 2024. Genetic mapping across autoimmune diseases reveals shared associations and mechanisms. *Nature Genetics* 56(5):838–845

[26] Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali TM. 2020. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods* 17(2):147–154

[27] Sasaki T, Bracero S, Keegan J, Chen L, Cao Y, et al. 2022. Longitudinal immune cell profiling in patients with early systemic lupus erythematosus. *Arthritis & Rheumatology* 74(11):1808–1821

[28] Lever J, Krzywinski M, Altman N. 2017. Points of significance: principal component analysis. *Nature Methods* 14(7):641–642

[29] Korem Y, Szekely P, Hart Y, Sheftel H, Hausser J, et al. 2015. Geometry of the gene expression space of individual cells. *PLoS Computational Biology* 11(7):e1004224

[30] Athey S, Imbens G. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences of the United States of America* 113(27):7353–7360

[31] Schiebinger G, Shu J, Tabaka M, Cleary B, Subramanian V, et al. 2019. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* 176(6):928–943. e22