Integrative, high-resolution analysis of single cell gene expression across experimental conditions with PARAFAC2-RISE

Authors: Andrew Ramirez[1], Brian T. Orcutt-Jahns[1], Sean Pascoe[1,2], Armaan Abraham[1], Breanna Remigio[3], Nathaniel Thomas[4], Aaron S. Meyer[1,5,6,*]

Affiliations:

[1] Department of Bioengineering, University of California, Los Angeles (UCLA), CA, USA
[2] Department of Molecular Biosciences, Northwestern University, Evanston, IL, USA
[3] Computational and Systems Biology, UCLA, CA, USA
[4] Department of Computer Science, UCLA, CA, USA
[5] Jonsson Comprehensive Cancer Center, UCLA, CA, USA
[6] Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, UCLA, CA, USA
[*] Corresponding author. Email: ameyer@asmlab.org

Keywords: PARAFAC2, tensor decomposition, single-cell analysis, scRNA-seq

Abstract: Effective and scalable exploration and analysis tools are vital for the extraction of insights from large-scale single-cell data. However, current techniques for modeling single-cell studies performed across experimental conditions (e.g., samples, perturbations, or patients) require restrictive assumptions, lack flexibility, or do not adequately deconvolute condition-to-condition variation from cell-to-cell variation. Here, we report that Reduction and Insight in Single-cell Exploration (RISE), an adaptation of the tensor decomposition method PARAFAC2, enables the dimensionality reduction and analysis of single-cell data across conditions. We demonstrate the benefits of RISE across two distinct examples of single-cell RNA-sequencing experiments of peripheral immune cells: pharmacologic drug perturbations and systemic lupus erythematosus (SLE) patient samples. RISE enables straightforward associations of gene variation patterns with specific patients or perturbations, while connecting each coordinated change to single cells without requiring cell type annotations. The theoretical grounding of RISE suggests a unified framework for many single-cell data modeling tasks. Thus, RISE provides an intuitive universal dimensionality reduction approach for multi-sample single-cell studies across diverse biological contexts.

Highlights:

- RISE enables tensor-based analysis of single-cell experiments across conditions.
- RISE separates condition-specific effects from cell-to-cell variation.
- RISE provides intuitive isolation of patterns into condition-, cell-, and gene-specific patterns.

INTRODUCTION

Single-cell measurements have revolutionized our ability to study the variation within and between heterogeneous cell populations[1,2]. As single-cell RNA-sequencing (scRNA-seq) and other high-dimensional single-cell measurement technologies have become increasingly accessible, these technologies have extended to studies including multiple experimental conditions, samples, or subjects. Multi-condition single-cell experiments evaluate how heterogeneous cell populations behave in response to different biological environments, such as how cells respond to drugs and gene knockouts or how cells differ across patients according to disease pathology[3]. There are substantial challenges in the analysis of these data, and appropriate analytical techniques are essential to learning from these studies.

The size and complexity of single-cell data present inherent challenges, including how to deconvolute biological variation within cell populations from variation arising from condition-specific differences. To address these challenges, studies have employed several computational approaches, such as pseudo-bulk techniques, wherein cells are coalesced into population averages, and differential expression is assessed for each population between conditions[4–6]. Other methods enable cell clustering, gene set enrichment, and pseudotime analyses to identify underlying trends in single-cell data[7–10]. Another source of variation includes technical/batch variations from sample processing. A few prominent batch integration techniques have emerged such as Seurat's canonical correlation analysis using linear combinations of features[11]. Others include Harmony’s iterative soft clustering to align similar cell types across batches, LIGER's integrative non-negative matrix factorization which identifies shared subspaces between datasets, and deep learning approaches like scVI, a stochastic optimization and deep neural network model[12–14]. However, batch correction methods introduce subjective bias and risk removal of meaningful biological differences or preserving technical noise.

Other methods for modeling such data focus solely on the analytical challenges of identifying biological variation across experimental conditions. One method, PopAlign, leverages a few high-level cell types to align a reference and a test sample with a Gaussian mixture model[15]. A plethora of complementary methods have emerged to detect biological differences between experimental treatments, including PhEMD and HSS-LDA[16–20]. Nevertheless, these methods either make restrictive assumptions about how the data is distributed[15,17,20]; do not directly isolate patterns of variation to specific conditions, cell populations, and genes[16–18]; or reduce the analysis of single-cell variability to a specific number of pre-determined cell types[19].

Many of these computational methods make use of matrix factorization as a key part of their workflow, as it allows large datasets to be summarized into interpretable and computationally tractable patterns that are each associated with unique sets of cells and genes. However, matrix factorization techniques fail to account for the inherently high-dimensional sources of variation in such studies (variation with regard to cells, genes, conditions) as they require that data be concatenated into a single two-dimensional data matrix where each row corresponds to a cell, regardless of its condition, and the column represents a gene that is measured[21]. When these multidimensional datasets are reduced to two dimensions, the distinction between cell- and condition-associated variations becomes conflated, obscuring the independent contributions of each dimension[22]. This conflation is a fundamental issue for analyzing multi-condition single-cell data and hinders the identification of biologically meaningful patterns.

One approach for analyzing measurements across several dimensions is rearranging the data into a tensor, or a multidimensional array, and using multilinear tensor decomposition techniques[23]. Canonical polyadic decomposition (CPD), one such tensor decomposition technique, approximates a dataset as the sum of the outer product of *n* vectors for an n-dimensional input, with each vector corresponding to a dimension of the data. CPD has been useful in bulk RNA-seq analysis measured across different tissues and patients; where the 3D structure of the data can be approximated by component patterns, each component having three vectors which explain how that pattern is associated with each gene, tissue, and patient, respectively[24]. Where these methods are appropriate, we and others have observed several benefits of a tensor decomposition-based analytical approach[24–26]; tensor decompositions can be more effective at removing noise, isolating distinct variation patterns, and imputing missing values compared to matrix-based techniques[24,27–29]. Most importantly, associating trends in the dataset to specific dimensions helps to interpret the results and therefore derive insights from the underlying experiments[24,30,31]. For example, tensor-cell2cell is a method that associates cell-cell communication signatures with specific experimental samples, using pseudobulk cell type profile information[28]. Dimension-specific associations also make these methods especially effective in data integration, combining datasets with shared experimental parameters[24,32–36].

Single cell experiments across conditions might be naturally organized into a three-dimensional tensor with axes representing the condition (e.g., samples, perturbations, or patients) in which the cells were collected, individual cells, and each gene measured. However, while conventional tensor decomposition techniques such as CPD assume alignment, or consistent correspondence across dimensions, single-cell technologies do not measure the same cell across conditions. One solution for decomposing data that lacks alignment along one dimension is an approach referred to as PARAFAC2[37]. PARAFAC2 derives a series of projection matrices for each condition to align the data across dimensions, enabling factorization of the data and thus comparative analysis across each sample. Originally developed to analyze data from chromatography experiments, where the elution time of molecules can vary from run to run, PARAFAC2 has been applied in various contexts, including electronic medical health records and spectroscopy experiments[38–40]. In addition, PARAFAC2 is the basis for Fast-Higashi, a computational technique that that addresses chromosomal organization in single cells, specifically addressing the challenge of comparing chromosomes of different sizes[41].

Here, we demonstrate that our modification of PARAFAC2, Reduction and Insight in Single-cell Exploration (RISE), provides an effective and general solution for integrating multi-condition single cell experiments by dissecting the variation found within and across conditions. RISE improves on the resolution of this analysis by directly modeling individual cells in each condition, enabling the identification of subpopulations of cells with distinct behavior that are missed by existing approaches. RISE is an unsupervised method that does not require external labeling of cells into cell types or categorization of the experimental conditions, thus reducing the number of potentially restrictive assumptions imposed on the data prior to modeling. Both benefits result in enhanced associations with disease status in observational studies, and nuanced insights about the signatures associated with patient features. RISE builds upon the well-developed theoretical foundations of PARAFAC2, with connections to other tensor decomposition algorithms, making RISE readily extensible to other challenges, such as integrating multi-omic or single-cell and bulk measurements. Based on these observations, we propose RISE as a broadly effective solution for analyzing single cell experiments across several samples or conditions.

RESULTS

RISE integrates single-cell measurements across conditions through the principle of parallel proportional profiles

Multi-condition single-cell data is typically analyzed by flattening/concatenation, where all single-cell experiments are combined into one large two-dimensional matrix of cells by genes (Fig. 1a). Consequently, once data reduction is performed on this matrix, the association of each component pattern with cells and conditions is confounded. Instead, one might imagine structuring this data as a condition by cell by gene tensor and then applying canonical polyadic decomposition (CPD) to create distinct component associations with each dimension of the single-cell dataset. While CPD is appropriate for data that can be grouped into corresponding cell types across samples (Fig. 1b), single cells exist across in a heterogenous manner across continuum of cell states in gene expression space rather than in discrete, rigid cell types[42].

One solution for analyzing such data is Reduction and Insight in Single-cell Exploration (RISE), our extension of PARFAC2 tailored for single cell biology. PARFAC2 builds upon Raymond Cattell's 1944 principle of parallel proportional profiles, which asserts that patterns across datasets can best be aligned by identifying a shared set of patterns (here, gene signatures) in the dataset that only vary in their proportional magnitude (Fig. 1c)[37,43–45]. In the context of scRNA-seq, this suggests identifying a common set of gene signatures across all cells and conditions and then assigning a magnitude that reflects the strength of that gene signature. This principle is particularly valuable for analysis of scRNA-seq because it allows for the interpretable characterization of cells with subtle variations across experimental samples, even among cells of different cell types. Unlike existing singe-cell integration that forcefully aligns transcriptional profiles between conditions, PARAFAC2 quantifies relative abundances of gene signatures while preserving condition-specific variation to avoid overcorrection of biological differences (Table S1).

PARAFAC2 introduces an additional constraint that enables solutions that are unique beyond trivial scaling and permutation differences, while avoiding restrictive assumptions such as non-negativity or a specific underlying cell distribution[15,43]. Most conservatively, PARAFAC2 assumes similar variable cross-product matrices between experimental conditions, although many successful applications violate this assumption[43]. With scRNA-seq, this translates to the expectation of a roughly similar gene-to-gene covariance structure between conditions[37]. Due to its applications outside the biological sciences, the scaling and convergence performance of PARAFAC2 has already been widely explored alongside additional constraints such as restriction to only non-negative factors[37,43,46–49]. We have developed RISE by optimizing PARAFAC2 specifically for scRNA-seq data by including refined parameter initialization, strategic parameter selection, use of Nesterov acceleration, and post-factor modifications (see in Methods).

As part of its fitting process, PARFAC2 defines a standardized reference frame across conditions. This is accomplished by Orthogonal Procrustes, an algorithm that uses a linear projection to summarize data into a predefined number of cell identities, here termed "eigen-states" (Fig. 1d). This results in one projection matrix per sample, which converts all the cells in that sample into the common eigen-states. Thus, an eigen-state represents a weighted collection of cells with a shared set of gene signatures. This identifies cells which share expression profiles present across experimental conditions. The projections only contain information about cell identity, not condition-to-condition variation; they represent which cells are relevant when encompassing an eigen-state, thereby mapping cells to an underlying common latent space. Thus, the projections provide a common basis to compare cells with similar or dissimilar gene expression behavior across all samples.

After projecting the data into an aligned tensor, the projected tensor is factored via CPD[23]. CPD applied to the eigen-state intermediate tensor results in the weighted sum of rank-one vector outer products for the genes, eigen-states, and conditions. Similarly to PCA, each component, or a set of three vectors, describes a specific variation pattern in the data. Each component is defined by three separate vectors defining the association of variation to specific genes, eigen-states and conditions. The combination of all vectors across one dimension are the factors and define how components vary across that one dimension. Because both the cell alignment and factorization are initially unknown, the two steps of cell alignment via the projections and CPD factorization are iteratively repeated until the solution converges (Fig. S1, Methods). Importantly, the CPD factorization enforces a low-rank structure, and the projection matrices map each experimental condition onto this shared low-dimensional space.

The PARFAC2 algorithm yields projections for each experimental condition and three factor matrices for the genes, eigen-states, and conditions. As part of RISE, we adjusted the scaling, sign interpretation, and order of the components and eigen-states to improve factor interpretation of the PARAFAC2 model (see in Methods). To determine which cells are highly associated with a specific component, the eigen-state factor matrix can be converted into a cell-specific factor matrix for each component (Fig. 1e). These “weighted projections” are derived by multiplying the eigen-state factor matrix and the projections across all conditions. We can then perform non-linear dimensionality reduction to visualize the projections in a low-dimensional embedding (Fig. 1e). This projection latent space describes cell-to-cell variation, separating any condition-specific effects into the factors. To illustrate which cells are associated with a given component, we can overlay the weighted projections on the projection embedding. Collectively, this enables RISE to effectively identify biological variation across experimental conditions and cellular states common across all conditions.

RISE separates cell-to-cell from condition-specific variation

To explore the utility of RISE, we first examined scRNA-seq data from Chen et al., a study wherein human peripheral blood mononuclear cells (hPBMCs) from one healthy donor were treated with a panel of 40 drugs from an FDA-approved immunomodulatory compound library, along with 6 control conditions (Fig. 2a)[15]. Previous analysis, using mixture modeling, isolated the transcriptional changes induced by each drug within two broad immune cell populations: T cells and monocytes (MO). Compounds were ranked based on their dissimilarity to the unperturbed controls to determine which experimental conditions most altered gene behavior. We surmised that RISE might improve on the resolution of the previous analysis by revealing smaller subpopulations with unique gene expression responses, as well as shared or distinct responses across drug perturbations. Because PCA represents the optimal linear dimensionality reduction technique and a baseline for variance explained, RISE was compared to PCA. RISE and PCA explained similar amounts of variance using the same number of components, despite RISE being a more mathematically constrained model (Fig. S2a)[37].

After fitting RISE to the Chen et al. data with 20 components, we visualized the cell-to-cell variation using a PaCMAP embedding of the RISE projections (Figs. 2b, d)[50]. We used PaCMAP, a non-linear embedding method, for cell embedding, as it has been shown to more reproducibly preserve local relationships and prevent distortion of global structures as compared to other methods like t-SNE and UMAP. As a basis of comparison, we also plotted the PaCMAP embedding of the PCA scores with the same number of components (Figs. 2c, e). Whereas the PCA-based embedding resulted in clusters defined by their drug perturbation, the RISE projections-based embeddings distributed cells from each condition across the embedding space (Fig. 2b, 2c). Similar results were observed for other compounds, suggesting that the projection matrices generated using RISE represent patterns that have been deconvolved from condition-specific varying, unlike the factors generated by PCA (Fig. S2b–e). We clarify that RISE is not a batch correction method, but rather an approach that maps cell-to-cell variation in the projections and defines condition-specific differences in the underlying factors, as supported by the above results. Manually annotating the cell types revealed that the RISE projection embeddings resolved many cell subpopulations despite condition-to-condition differences (Fig. 2d, e). To quantify these effects, we observed that the PCA standard embedding resulted in poorer integration accuracy in both batch integration and conservation of biological variance in immune cell types, most likely due to clustering cells based on perturbation-specific gene expression responses (Fig. S2f)[51]. As expected, given this experiment involves short-term treatment of PBMCs from one donor, abundant cell types were similarly represented across all conditions (Fig. S3a-g).

Like PCA, RISE results in a series of component patterns, but with vector loadings representing the association of the pattern along the three dimensions of the data: drug perturbations (Fig. 2f), eigen-states (Fig. 2g), and genes (Fig. 2h). Each component pattern can be traced across the dimensions in the dataset and illustrates a distinct biological trend. A small fraction of genes was highly associated (i.e., highly weighted) with each component, consistent with our expectation that the effects of immunologic drugs on various cell types are defined through specific gene signatures (Fig. 2h).

To determine whether RISE condition and gene factors trends were robust to changes in cell number, we applied RISE to subsets of the drug perturbation panel. Using the Factor Match Score (FMS) to quantitatively measure the similarity of the factor matrices, we found that the results were essentially unchanged with the dataset reduced to as much as half its original size (Fig. S4a)[49]. This highlights the fact that the RISE eigen-states indeed correspond to robust signatures in the underlying data, as they are reproducible when only a representatively sampled subset of the cells are present. Using a bootstrapping strategy, we similarly observed consistent factor results at varying number of components, though a decreasing trend in the FMS at the highest number of components, as expected (Fig. S4b). These results indicated that RISE provides reproducible results and that we were not overfitting these data. Having confirmed the computational suitability of RISE for modeling our data, we next wondered what biological insights RISE can identify.

RISE identifies cell-to-cell, condition-specific, and context-dependent gene expression variation

Having assessed the reliability of the RISE factorization, we next inspected the patterns of gene expression variation that were identified (Fig. 2f–h). To interpret the biological significance of a component, one can inspect the associations across specific conditions, genes, and cells. Cell types were manually annotated using cell type marker genes at higher resolution to improve interpretation of the RISE components (Fig. 3a, Methods). One of the first advantages of RISE is the ability for components to identify common cells with similar gene expression profiles across experimental conditions. For instance, we plotted the component 15 weighted projections scores of individual cells on the PaCMAP embedding and found that plasmacytoid dendritic cells (pDCs) were selectively and highly weighted (Fig. 3e). FXYD2, SERPINF1, and RARRES2 are genes which have been identified as selectively expressed in pDCs in the literature and Human Protein Atlas[52–54]. These three genes were identified as the most strongly weighted in component 15, and further examination of these genes confirmed their specificity in pDCs (Fig. 3b). Examining the conditions factor, component 15 was largely invariant in its association with drugs, suggesting this component represents a gene expression module that is equally present across all conditions (Fig. 3c). Taking these findings together, we can conclude that component 15 represents the presence and unique gene expression of pDCs across all drug treatment conditions. Only 5 pDCs were found within each condition on average, demonstrating RISE's sensitivity to low-abundance cell populations, as well as its ability to reveal meaningful variation outside of contrasts between cases and controls (Fig. S5m). In addition to pDCs, RISE identified that component 12 was strongly associated with the amount of B cells in each drug perturbation experiment based on the genes identified by this component, including MS4A1A and CD79A (Fig. S6l, S7a, Table S2a). To simulate data where a cell population is differentially abundant in an experimental condition, we synthetically reduced or removed B cells from a drug perturbation. The resulting RISE factorization showed a reduced weight for component 10 in the altered condition, demonstrating how RISE identifies experimental conditions with differentially abundant cell populations within a condition (Fig. S7b, c).

In addition to representing the abundance of specific cell populations, RISE also identified patterns associated with specific drug perturbations. For example, component 19 was strongly and uniquely associated with alprostadil, a prostaglandin inducer of vasodilation and angiogenesis (Fig. 3c) [55]. Component 19 most strongly weighted the DC population (Fig. 3h), and the component-associated genes included various ones relating to angiogenesis such as THBS1, VMO1, CXCL5, EREG and VEGFA (Table S2b). VEGFA and EREG have been previously confirmed to be upregulated in response to prostaglandin stimulation[56,57]. Alprostadil-treated DCs expressed the most EREG and THBS1, and were the only DCs to co-express these, demonstrating RISE's ability to identify coincident patterns of gene expression within individual populations (Fig. 3i, 3j). Thus, component 19 represents an angiogenic gene module expressed in DCs that is uniquely induced by alprostadil and illustrates RISE's ability to model condition- and cell population-specific expression patterns.

RISE's tensor-based structure can also identify expression modules shared across several drug perturbations. For example, component 20 has a strong relationship with all glucocorticoids in the dataset. Glucocorticoids are steroid hormones commonly used to treat inflammation, autoimmune diseases, and cancer and are known to induce broad transcriptomic effects across immune cell populations[15]. The variance encoded by component 20 is strongly represented in the monocyte population, with distinct effects by monocyte subtype (Fig. 3l). Here, intermediate MO and myeloid suppressors had positive weights, while classical monocytes (MOs) had negative weights, suggesting that the gene expression and abundance information encoded by component 20 is divergent across these subpopulations. Cells in the myeloid suppressor cluster were more abundant in the glucocorticoid-based experiments, mirroring the cell specificity of component 20 (Fig. 3m). Indeed, when examining the cellular composition of samples treated with glucocorticoids, we observed a substantial decrease in MOs, and an increase in myeloid suppressor cells (Fig. 3n). The distributions of all other cell types remained consistent across conditions, suggesting that our projections-based embedding was not simply being informed by drug perturbation, as the PCA-based embedding was. The component 20 weight positively correlated with the proportion of cells labeled as myeloid suppressors (Fig. 3o). This suggests that component 20 captures glucocorticoid-induced expansion of immunosuppressive phenotypes within the myeloid compartment[58–60].

Although component 9 was also glucocorticoid-specific, it isolated distinct gene signatures and cell associations as compared to component 20 (Fig. 3c/p, Table S2a, S2b). Component 20 described glucocorticoid-induced CD163 and MS4A6A expression across MOs and other cell types, as seen in earlier studies (Fig. 3b, q, S7) [61]. CD163 upregulation in intermediate MOs correlated with MS4A6A expression in myeloid suppressor cells (Fig. 3r), suggesting coordination between these cell types and demonstrating RISE's ability to identify shared expression changes between populations. These observations indicated that RISE could identify both condition-specific expression patterns and cellular abundance differences, as demonstrated by its myeloid subtype analysis. However, because RISE simply identifies gene expression signatures that vary in abundance, each component must be examined to determine whether it represents cellular abundance of a cell type or cellular expression differences. Overall, RISE was able to effectively separate distinct and highly interpretable patterns of gene expression variation, whether cell type-specific (Fig. 3e/f), condition- and cell type-specific (Fig. 3h/i), or even conditions with shared alterations among subpopulations of cells (Fig. 3o–q). Here, we only highlight a select few patterns; we also found that the other components identified by RISE identify relevant biological insights as well, each of which is explored in Table S3.

RISE improves the resolution and accuracy of transcriptomic associations with systemic lupus erythematosus

With an understanding of the benefit RISE can provide within perturbational studies, we next examined its potential to separate cell-to-cell from patient-to-patient variation and to generate interpretable patterns which are predictive of patient status within an observational cohort study. To do so, we focused on a recent study that characterized hPBMCs from a cohort of 162 systemic lupus erythematosus (SLE) patients and 99 healthy controls by scRNA-seq[62]. Multiple scRNA-seq samples were collected from some patients, resulting in 209 SLE and 145 healthy samples. Samples varied widely in overall cell number (Fig. S8a). The previous analysis aggregated single-cell gene expression into pseudobulk profiles for high-resolution cell types and performed differential expression analysis. Approximately 300 differentially expressed genes for groups of immune cells (e.g., all myeloid cells or all immune cells) were used as features for a model distinguishing SLE from healthy samples[62].

We first investigated RISE's general explanatory and associative power. RISE explained an increasing fraction of the dataset variance with more components (Fig. S8b). To select an appropriate number of components, we used a logistic regression classifier to associate the patient (i.e., condition) factors with SLE status. Cross-validation was performed identically to the previous study, using batch-specific training and testing; briefly, we trained the logistic regression model on one of the four processing batches of data and then used the patient sample factors to predict the status of the left-out batches[62]. Increasing the component number improved training accuracy for SLE up to 30 components (Fig. 4a, S8c). Bootstrapping the samples or using leave-one-batch-out cross-validation did not strongly alter these observations (Fig. S8d-f). A 30-component RISE model was predictive of SLE status using the same cross validation as the previous authors and resulted in an AUC ROC of 0.94 (Fig. 4b). This approach outperformed both the previously used pseudobulk prediction analysis, and our analysis using pseudobulk gene expression and cell type composition as predictors to classify SLE status. The 30 components were variously associated with the patients' samples, eigen-states, and genes (Fig. S8g). Examining the cell-to-cell variation within the projections via the PaCMAP embedding, we observed effective separation of the expected immune cell types in hPBMCs, both at lower and higher resolution (Fig. S8h, 4c). In addition, we compared RISE's scalability and computational efficiency and found that it compares favorably to other single-cell analysis tools (Fig. S9a-d).

An L1-penalized logistic regression model identified components relevant for differentiating SLE sample status (Fig. S9f). However, L1 regularization eliminated components that could possibly differentiate SLE status. We therefore also assessed the classification accuracy using all pairwise combinations of components (Fig. 4d). Notably, components 14, 22, 27, and 28 led to accurate classification in reduced models. We again traced the components across the cell and gene dimensions to interpret the significance of each RISE pattern (Fig. 4e, Table S4a, b).

We first investigated component 14 as it was most strongly SLE-associated. Component 14 selectively associated with Naïve CD4+ and CD8+ T cells, as well as a subcategory of proliferating T cells (Prolif) (Fig. 4f). The sample associations of this component directly correlated with the abundance of these annotated lymphocyte populations (Fig. 4g). Consistent with explaining the abundance of these cell populations, the genes associated with component 14, including CCR7, LEF1, and PRKCQ-AS1 (Fig. 4e), are crucial for early T cell development, as seen similarly by the previous analysis of the data[63,64]. Gene set enrichment analysis (GSEA) of the 30 most weighted genes for component 14 confirmed an association with T cell activation and lymphocyte differentiation (Fig. 4h, Table S4b).

Component 22 was also associated with SLE status and corresponded to the myeloid populations of CM, nCM and cDC1 (Fig. 4i). Similarly to the T cell development pattern, component 22 correlated with the cell abundance of these specific myeloid populations (Fig. 4j). This component was associated with interferon (IFN)-induced genes, including IFI44, MX1, IFITM3, APOBEC3A, and ISG15 (Fig, 4e). Other components overlap in these IFN genes, such as component 6, but highly weight distinct cell populations (Table S4a, b). The IFN-stimulated cell types identified by component 22 uniquely express IFITM3 and APOBEC3A in combination (Fig. 4k). Thus, RISE recovered several known alterations associated with SLE[62,65].

RISE identifies novel single-cell patterns associated with SLE

We next explored whether other components might yield novel insights into SLE-associated mechanisms (Fig. 5a). We first inspected component 4 due to its association across several cell types, including GZMH+ CD8+ T cells and several NK populations (Fig. 5b). SPON2, FGFBP2, GZMB, PRF1 and GZMH were among the highly weighted genes associated with component 4 (Fig. 5a, S11). Many of the genes identified by this pattern also overlap with the genes previously used to define a cytotoxic signature for GZMH+ CD8+ T cells[62]. Using the same cytotoxic signature to quantify this gene module as the original authors, we found that other immune cell types also expressed this signature, including NK Bright, NK Dim, and T8 GZMK+ CD8+ T cells, as identified by component 4 (Fig. 5c). This finding demonstrated that the RISE can uncover gene expression changes that are coordinated between cell populations which would have been missed in per-population analyses[66,67].

In addition to identifying coordinated changes across cell populations, we found that RISE identified gene modules in subpopulations of cells that would be missed by a pseudobulk approach. For example, component 27 was highly SLE associated and specifically weighted a small subcluster within the CM population comprising 1% of the total population of cells per patient sample (Fig. 5d). The most highly weighted gene in this component, RETN, was found to be uniquely expressed in cells identified by high component 27 weights (Fig. 5a, 5d, 5f). In addition to RETN, the genes S100A8, S100A12, and S100A9 were highly weighted, which others have identified as relevant to IL-6 signaling and activation of toll-like receptors 4 and 7 in SLE patients, prolonging inflammation (Fig. 5a, S11)[68–71]. The component 27-high CM subpopulation is distinct in its shared expression of RETN, S100A8, S100A12, and S100A9 as compared to all other cells (Fig. 5g). RISE revealed the increase in gene expression of the inflammatory module within a subset of classical monocytes, a nuance which would been obscured in a pseudobulk analysis that examined all myeloid cells.

Furthermore, RISE can identify subpopulations of traditional cell types between cell types with similar expression signatures. Notably, component 28, which also had high association with SLE status (Fig 4d), is highly specific to CM and CD4+ effector memory T cell (TEM) subpopulations (Fig. 5h). The component 28 sample weightings were most correlated with the cell abundance of the CM subpopulation and component 28 identified many IFN genes including IFI27, IFI6, ISG15, IFITM3 and IFI44L (Fig 5c). Although some of these genes are highly weighted in other components like in component 22, component 28 strongly weights IFI27 which is specific to these CM and CD4+ TEM subpopulations (Fig. 5i, S11). As expected, the subclusters for both CM and CD4+ EM have a higher average expression of IFI27 in SLE patient samples (Fig. 5j). Thus, one can conclude there are subclusters of both CM and CD4+ effector memory T cells that share this unique expression of IFI27, along with other type 1 interferon regulated genes. Indeed, IFI27 has been suggested as a potential biomarker for SLE in CD4+ T cells and monocytes[65,72–74]. However, our finding hints that not all cells categorized as CD4+ EM and CM contribute to this unique expression gene module.

DISCUSSION

Here, we demonstrated that RISE improves the investigation and understanding of multi-condition single cell experiments. By aligning cells between conditions during the model fitting process, single-cell data can be directly factored using tensor decomposition, thus preserving the experimental structure more faithfully than matrix factorization methods would. Because RISE builds upon PARAFAC2, an unsupervised data reduction method, no prior information is required about cell identity or the types of experimental conditions, making RISE unbiased as compared to other models that utilize a reference atlas[75,76]. In both the Chen et al. and Perez et al. studies above, RISE identified transcriptional patterns that were shared across conditions as well as those that were specific to a subset of conditions. These patterns of expression were quantitatively linked to individual cells making them highly interpretable. Overall, these results show how a tensor-based approach to single-cell data leads to unique insights that cannot be identified by dimensionality reduction in matrix form or standard pseudobulk analyses, both which obscure biological relationships across conditions. RISE eliminates the need to define the difference between technical and biological variation by characterizing patterns of variation across single cells and experimental conditions in a data-driven manner. RISE thus serves as an alternative single-cell analysis tool that effectively isolates and deconvolutes variation arising from cells, genes and experimental samples, without the need to perform batch correction, or to organize data in a way that obfuscates patterns across dimensions (Fig. 6).

The advantage of analyzing single cell experiments across conditions with RISE was first demonstrated for the exploratory analysis of drug perturbations. While earlier analysis demonstrated the ability to identify perturbed populations at a high level, distinguishing T cells and monocytes, the RISE components provide quantitative information about perturbations integratively and at high resolution. Despite not requiring cell annotations, RISE isolated rare subpopulations, such as pDCs, and their canonical markers including FXYD2 and RARRES2/chemerin[77]. It also distinguished several qualitatively distinct glucocorticoid responses, including conversion of cell identity to myeloid derived suppressor cells and variation in the types of responses across monocyte subsets[78]. While prior analysis localized Alprostadil effects to monocytes overall, RISE demonstrated that DCs were most affected by Alprostadil treatment and identified an associated pro-angiogenesis expression signature.

The strengths of RISE were especially evident in more complex single-cell transcriptomic data from SLE patients. While earlier analysis of this data relied on pseudobulk methods, we demonstrated that by directly modeling the single cell data, we were able to improve our identification of patterns which were predictive of SLE status. RISE confirmed known alterations in naïve T cell and monocyte abundances, while also recovering differences in CD8+ T cell maturation, proliferating T cells, and cDC1 cell abundances described by others[67,79–81]. Our analysis also uncovered a coordinated cytotoxic and exhaustion-like GZMH+ expression program that extends beyond CD8+ T cells to NK cells, B cells, and CD4+ T effector memory cells, the last of which have been recently found to be associated with SLE disease progression[66,82]. Additionally, RISE highlighted gene expression modules in a small subset of monocytes with unique expression of RETN and S100A genes. Other studies have suggested S100A8 and S100A9 as potential biomarkers for SLE and this analysis points to subpopulations of CM as responsible for this observed pro-inflammatory behavior in SLE[68,69]. Similarly, RISE identified upregulated coordinated changes of IFI27, another putative SLE biomarker, in specific subsets of monocytes and CD4+ T effector memory cells[83,84]. Many SLE studies have shown the autoimmune disease is characterized overall by a type 1 IFN-regulated response, but RISE delineated distinct IFN signatures according to their gene patterns and cellular localization[85,86].

Despite the popularity of pseudobulk analysis, these results demonstrate several of the limitations of this approach. One must choose a level of detail at which to annotate the cells, with the ultimate results and even basic features such as statistical power affected by the number of populations the cells are split into. These results are reliant on accurate cell annotations, and small fractions of inconsistently annotated cells could lead to artifacts in differential gene expression analysis. Clustering algorithms are often employed to form populations, which are then annotated. However, there are no guarantees that clustering populations align with their biologically significant counterparts, especially in the presence of technical artifacts between batches of cells[87]. Furthermore, particularly in smaller studies, rare cell-types are often unidentifiable and clustered among higher-abundance cell types, obscuring their potentially powerful contributions to pathology or condition-specific variation. Finally, cells exist in a continuum of potentially overlapping identities, which are impossible to completely capture as distinct clusters—for instance, naïve, effector, and memory populations exist across CD4, CD8, and NK cell populations. Consequently, it is not possible to form accurate and discrete classes of cells. Instead, RISE offers an alternative perspective of cell “eigen-states”, which are multivariate and capture these distinct types of identities.

Both the strengths and limitations of RISE arise from the fact that it is a linear dimensionality reduction method. The method does not attempt to explicitly identify or correct for batch effects or other technical artifacts. Indeed, while there have been many attempts to understand and correct for these issues, there is a lack of agreement as to the extent of technical artifacts within scRNA-seq measurements, and many study designs do not enable a definitive line to be drawn between biological or technical variation[51,88].

Instead, RISE can be used to identify patterns that are associated with batches; for example, by training a model on the condition factor matrix that predicts what batch a condition was processed in, we can understand which components reflect batch specific variation. By identifying component patterns, these gene expression alterations can be isolated and inspected. RISE requires the number of components as a model input; similarly to PCA, defining the optimal number of components can be difficult, with different heuristics being potentially appropriate depending on the goal of the modeling effort. In cases where one's goal is to identify patterns predictive of disease, cross-validation using the factors as features may be appropriate. When there is not an association prediction or other task with which to evaluate the resulting factors, other metrics such as FMS or variance explained may aid in the task of selecting the size of the RISE model[37,89].

An important strength of RISE is that it provides a foundational framework that can be expanded across various tasks. Additional constraints on the factors, such as nonnegativity, can enhance interpretation of the results[90]. The RISE algorithm can be altered to be a supervised factorization approach, leveraging information about each sample to better identify patterns that correspond to features on the cell or condition/sample scale[91]. RISE is well-suited to analyze single-omic analysis, but can be broadened to perform coupled decomposition for multi-omic data, such as combining transcriptomics with proteomics, to identify associative patterns across conditions[92]. With modest extension, such as constraints to account for spatial localization, RISE could be extended for identifying gene expression patterns in spatial measurements[49,93].

In summary, we find that RISE is highly effective for modeling and exploring multi-condition single cell studies in both perturbational experiments and in an autoimmune cohort. RISE localizes variation to specific samples, genes, and populations of cells while enabling interpretation beyond what is possible with non-linear approaches such as autoencoders[14]. As the scale, capabilities, and questions asked using single cell techniques expand, fundamentally new approaches such as RISE that are scalable, easily interpretable, and unsupervised will be needed to understand how single cells coordinate their function within and across tissues[35].

Acknowledgements: This work was supported by NIH U19-AI172713 to A.S.M., an Emerging Leader Award from the Mark Foundation for Cancer Research, and an NSF Graduate Research Fellowship to A.R.

Author Contributions: A.S.M. conceived of the study. A.R., B.O.-J., S.P., B.R., N.T., and A.S.M. performed the computational analysis. A.R., B.O.-J., S.P., B.R., N.T., and A.S.M helped to analyze the data. A.R., B.O.-J, S.P., A.A., and A.S.M. contributed to writing the paper.

Declaration of Interests: The authors declare that they have no competing interests.

Supplemental Information: Supplemental Information is attached including plots for Figures S1–10 and descriptions for Figures S1–10 and Table S1–5. Content of Table S1–5 will be attached as Excel files.

STAR Methods

Resource Availability

Lead Contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Aaron Meyer (ameyer@asmlab.org).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

- All analysis was implemented in Python v3.12 and can be found at https://github.com/meyer-lab/scCP, release 1.0.
- The implementation of RISE is available at https://github.com/meyer-lab/parafac2. The single-cell gene expression was obtained from Figshare (https://doi.org/10.6084/m9.figshare.11837097) and via GEO accession number GSE174188.
- Any additional information required to reproduce this work is available from the Lead Contact.

Figure Captions

Figure 1. RISE enables the interpretable integration of single-cell measurements across conditions. (a) Traditional data reduction methods are applied to flattened multi-condition data in matrix form. **(b)** Instead, tensor decomposition of aligned multi-condition data results in loadings describing the component association with each dimension. **(c)** PARAFAC2 satisfies the principle of parallel proportional profiles which states that latent variable models can be most effectively aligned by identifying common underlying patterns (here, gene signatures) in the dataset that only vary in their proportional magnitude within each condition or sample. **(d)** Multi-condition single-cell experiments can be naturally organized into a staggered three-dimensional tensor; PARAFAC2 uses Orthogonal Procrustes to project cells onto common cell eigen-states and tensor decomposition (CANDECOMP/PARAFAC) to factor this aligned data. **(e)** The projections for each experimental condition multiplied by the eigen-state factors provide an association of each cell and component pattern. The projections alone also summarize the cell-to-cell variation independent of condition-specific effects.

Figure 2. RISE defines cell-to-cell and drug-specific variation in a drug perturbation scRNA-seq experiment. (a) hPBMCs were treated with 40 FDA-approved drugs, including glucocorticoids, prostaglandins, mTOR inhibitors, and calcineurin inhibitors, alongside six controls. **(b–c)** PaCMAP embedding of the (b) RISE projections and (c) PCA-based dimensionality reduction, with cells from the alprostadil treatment condition highlighted. **(d–e)** PaCMAP of the (d) RISE projections and (e) PCA-based dimensionality reduction, with cells colored by their cell type annotations. **(f–g)** The (f) drug condition, (g) eigen-state, and (h) gene factors for each component. The genes factor matrix is filtered for genes having an effect magnitude of 0.8 or more within one of the components. The conditions factors are non-negative and have been log transformed to better show weights relative to the control conditions.

Figure 3. RISE analysis of drug perturbations reveals both broad and highly localized gene expression responses. (a) PaCMAP of the RISE projections, with cells colored by their high-resolution cell type annotations. **(b, c)** The (b) gene and (c) condition factors matrices with select genes and conditions highlighted. **(d, g, k)** Schematics of the biological significance of components (d) 15, (g) 19, and (k) 20 describing each component's association with genes, eigen-states and conditions. **(e, h, l, p)** PaCMAP of the RISE projections, with cells colored by the weighted projections (WP) for components (e) 15, (h) 19, (l) 20, and (p) 9. **(f)** Per-condition average expression of FXYD2, RARRES2 and SERPINF1, stratified by cell type. **(i)** Average gene expression of EREG across drug conditions for low-resolution cell types. **(j)** Expression of THBS1 and EREG in individual dendritic cells, with alprostadil-treated cells labeled. **(m)** PaCMAP of all cells colored by whether they were treated with glucocorticoids. **(n)** Proportion of cells classified as each cell type for each condition, stratified by glucocorticoids versus all other drugs. **(o)** Percentage of cells classified as classical myeloid suppressors for each condition and their corresponding component 20 weight. **(q)** Average expression of MS4A6A within each cell type stratified by whether cells were treated with glucocorticoids or all other drugs. **(r)** Average CD163 expression in intermediate monocytes versus average MS4A6A expression in myeloid suppressors for each condition.

Figure 4. RISE improves the resolution and accuracy of transcriptomic associations with systemic lupus erythematosus. (a) ROC AUC of a logistic regression classifier, using the patient factors, predicting SLE status. RISE rank was varied between each model. **(b)** ROC curve of a logistic regression classifier, using the patient factors, predicting SLE status with a 30-component RISE model. Cross-validation was performed by leave-one-processing-batch out. **(c)** PaCMAP of the RISE projections, with cells colored by their higher resolution cell type annotations. **(d)** SLE prediction accuracy of an unpenalized logistic regression model trained on pairs of components. **(e)** The gene factors matrix using 30 components, with genes from components 14 and 22 highlighted. **(f, i)** PaCMAP of the RISE projections, with cells colored by the weighted projections of components (f) 14 and (i) 22. **(g, j)** Pearson correlation coefficient based on the relationship of cell type count per patient against corresponding patient sample weightings of components (g) 14 and (i) 22. **(h)** Gene set enrichment analysis (GSEA) biological process results for component 14 gene module (top 30 positively weighted genes), where dot color represents the number of genes overlapping with the gene ontology. **(k)** Average expression of IFITM3 and APOBEC3A for low-resolution cell types in healthy and SLE samples.

Figure 5. RISE identifies high-resolution subpopulation and gene signatures associated with SLE. (a) The genes factors matrix with select genes highlighted. **(b, d, g)** PaCMAP of the RISE projections, with cells colored by the weighted projections of components (b) 4, (d) 27, and (g) 28. **(c)** Cytotoxic gene score stratified by each cell type. **(d)** Normalized RETN expression of single cells overlayed on PaCMAP embedding **(f, j)** Average (g) RETN, S100A8,

S100A12, S100A9, and (k) IFI27 gene expression classified by SLE status and subcluster. **(h)** Pearson correlation coefficient based on the relationship of subcluster cell count and condition sample weightings per patient for component 28. **(i)** Correlation between component 22 and 28 gene weightings, with IFI27 highlighted.

Figure 6. RISE presents an innovative, distinct single-cell analysis method to map interconnected variance in gene modules and cellular populations across experimental samples, providing a comprehensive framework for understanding complex cellular heterogeneity.

Method Details

PARAFAC, PARFAC2, and RISE

Our implementation and interpretation framework around the PARAFAC2 tensor decomposition algorithm, which has been optimized for its use towards scRNA-seq, is coined Reduction and Insight in Single-cell Exploration (RISE). We use RISE to refer to our adaptation of PARAFAC2 which includes parameter initialization scheme for PARAFAC2, parameter selection of PARAFAC2, use of Nesterov acceleration for optimization, and interpretation utilities of the PARAFAC2 factors.

PARAFAC

As PARAFAC and PARAFAC2 serve as the core modeling approach behind RISE, we will first provide a mathematical explanation of PARAFAC. PARAFAC is a generalization of principal components analysis, which primarily operates on second-order tensors (matrices). Despite its applicability to arbitrary-order tensors, we will explain the usage of PARAFAC to third-order tensors, as this is most relevant to our subsequent PARAFAC2 explanation. Given a third-order tensor, $\boldsymbol{X}$, of shape $I \times J \times K$, PARAFAC seeks to approximate $\boldsymbol{X}$ as a sum of component rank-one tensors. This set of component rank-one tensors are collectively represented in factor matrices: A of shape $I \times R$, B of shape $J \times R$, and C of shape $K \times R$, where R is the number of components into which $\boldsymbol{X}$ is decomposed. Specifically, PARAFAC seeks

$$\underset{A,B,C}{\operatorname{argmin}} \sum_{i=1}^{I} \| X_i - B diag(a_i) C^T \|^2$$

where $X_{\boldsymbol{i}}$ is the i^{th} slice of $\boldsymbol{X}$ along the first mode, and $diag(a_i)$ is the diagonal matrix whose nonzero entries are equal to the i^{th} row of A.

PARAFAC2

In single-cell transcriptomics, a different set of cells is profiled for each condition, and the number of cells profiled in each condition may vary. This precludes the direct arrangement of these measurements into a single tensor, and prevents the use of PARAFAC and many other existing tensor-based decomposition methods[94]. This data would instead be better represented as a length-I list, $\boldsymbol{X_{sc}}$, of second-order tensors, $X_{sc,i} \in \mathbb{R}^{\mathrm{J}_{\mathrm{sc},i} \times \mathrm{K}}, \ \mathrm{i} = 1, 2, \ldots, \mathrm{I}$, where I is the number of conditions, $J_{sc,i}$ is the number of cells profiled in condition i, and K is the number of genes. PARAFAC2 explicitly handles data of this form, where one axis (in this case the cell axis) varies across another (here, the condition axis). We use the direct fitting algorithm originally proposed by Kiers et al. to fit the PARAFAC2 model[37].

In fitting the PARAFAC2 model, we seek

$$\underset{A,B,C,\boldsymbol{P}}{\operatorname{argmin}} \sum_{i=1}^{I} \left\| X_{sc,i} - P_i B diag(a_i) C^T \right\|^2 \text{ subject to } {P_i}^T P_i = \boldsymbol{I} \ \forall \ \mathrm{i}$$

where $\boldsymbol{P} = P_i \in \mathbb{R}^{\mathrm{J}_{\mathrm{sc},i} \times \mathrm{R}}, \ \mathrm{i} = 1, 2, \ldots, \mathrm{I}$ are the projection matrices, where R is the rank of the PARAFAC2 decomposition; $diag(a_i)$ is the diagonal matrix whose nonzero entries are equal to the i^{th} row of $A \in \mathbb{R}^{I \times R}$, the condition factor matrix; $B \in \mathbb{R}^{R \times R}$ is the cell eigen-state factor matrix; $C \in \mathbb{R}^{K \times R}$ is the gene factor matrix; and $\boldsymbol{I}$ is the identity matrix[37].

Comparing the objective functions for PARAFAC and PARAFAC2, one can see that the difference between the two models is simply the addition of the orthonormal projection matrices $\boldsymbol{P}$ in PARAFAC2, which accounts for data that is unaligned in one dimension, such as scRNA-seq.

We fit PARAFAC2 using alternating least-squares, whereby $\boldsymbol{P}$ and the factor matrices, A, B, and C, are alternately fit. First, the factor matrices are updated as

$$\underset{A,B,C}{\operatorname{argmin}} \sum_{i=1}^{I} \left\| X_{sc,i} - P_i B diag(a_i) C^T \right\|^2$$

$$= \underset{A,B,C}{\operatorname{argmin}} \sum_{i=1}^{I} \left\| P_i^T X_{sc,i} - B diag(a_i) C^T \right\|^2$$

which is equivalent to running PARAFAC[94] on the single, aligned, third-order tensor whose i^th slice along the first mode is equal to $P_i^T X_{sc,i}$[37]. Then, the projection matrices, $\boldsymbol{P}$, are updated by solving the Orthogonal Procrustes problem. Given matrices $\mathrm{X}, \mathrm{Y} \in \mathbb{R}^{\mathrm{n}\times\mathrm{m}}$, the Orthogonal Procrustes problem seeks an orthogonal matrix $Q \in \mathbb{R}^{n\times n}$ that minimizes $\|XQ - Y\|_{\mathrm{F}}^2$. The solution is given by $\mathrm{Q} = \mathrm{UV^T}$, where $\mathrm{U\Sigma V^T}$is the singular value decomposition of $\mathrm{Y^T X}$. For a low-rank decomposition, we update each $\mathbf{P}_{\boldsymbol{i}} \in \boldsymbol{P}$ as the product of the first R left and right singular vectors from the SVD of $\mathrm{B}diag(a_i)C^T X_{sc,i}^T$[37]. This process of updating the factor matrices using PARAFAC and the projections via Orthogonal Procrustes is repeated until a specific number of iterations elapses or the difference in R2X between iterations reaches a threshold.

RISE

Parameter Initialization

Parameter initialization was performed by using the SVD of the matrix-flattened data, as previously described[37]. Each of $X_{sc,i}$ was concatenated along the cell dimension to form a cell by gene matrix containing cells from every condition. Randomized SVD[95] was performed on this matrix and the first R right singular vectors were used to initialize B. A was filled with ones and C was initialized to the identity matrix.

Optimization Technique

We used the TensorLy implementation of PARAFAC and alternating least squares (ALS) to solve for each factor, wherein each mode is iteratively solved using least squares[96]. We used 20 iterations of PARAFAC per PARAFAC2 iteration. We terminated the PARAFAC2 fitting algorithm once either 200 iterations had elapsed or the difference in R2X between iterations was less than 10^{-6}.

To speed the fitting process, we use all-at-once Nesterov-like acceleration for fitting the model as previously described[97], with extrapolation step size determination parameters of $\gamma = 1.1, \bar{\gamma} = 1.03, \eta = 1.5, \beta_0 = 0.05, \overline{\beta_0} = 1$.

Post-Factorization Alignment

Like PCA, CP-based decomposition component patterns are made up of the product of vector weights for each dimension in the data. CP-based decompositions generally provide a unique solution, aside from three indeterminacies: the sign of the factors, scaling between factors, and ordering of the components. We therefore applied rules to yield consistent factorization results, helping with the interpretation of the results. To address the scaling indeterminacy, we normalized each factor matrix to have components of unit scale. The sign indeterminacy was partially resolved by making the conditions factors positive.

PARAFAC2 factors are order indeterminant, where the order of each component can be arbitrarily arranged. $\boldsymbol{A}$, the conditions factor, was used to arrange components according to their Gini coefficient of association across the conditions, so that the first components were generally found to be associated with all the conditions, while the later components were generally condition specific.

Reconstruction Error

We used the metric reconstruction error (R2X), or the percentage of variance explained by the model, to assess the model fit. First, the Frobenius norm squared of the difference is calculated between the data (X_{data}) and the reconstruction based of the factors and projections (X_{Pf2}). This difference is normalized by the Frobenius norm of the data and subtracted from 1 to result in a range from 0 to 1, where 1 represents 100% variance explained.

$$R2X = 1 - \frac{||X_{data} - X_{Pf2}||^2}{||X_{data}||^2}$$

Factor Match Score

The factor match score (FMS) measures the similarity between two sets of factors[49]. Two FMS strategies were implemented. First, to determine how the size of the data influences the factors, the FMS was compared between RISE factorizations with the same number of components on the entire dataset and with various fractions of the cells removed. Second, to determine the stability of the factorization at varying numbers of components, cells were bootstrapped to create a dataset of the same size. The FMS was then calculated for the original dataset and the reconstructed dataset at various components. Thus, after 2 unique RISE factorizations are applied in both cases, the FMS compares the conditions (A) and genes (B) factor matrix of the original factorization to the reconstructed factor matrices, $\hat{A}$ and $\hat{B}$. The eigen-states factors vary in what cells they represent, so they are not included in calculating the FMS. An FMS of 1 indicates complete similarity. A linear sum assignment is also applied because a decomposition can have the same components, but in a different order.

$$FMS = \sum_{r=1}^{R} \left(1 - \frac{w_r \widehat{w_r}}{max(w_r \widehat{w_r})}\right) \cdot \frac{A_r^T \widehat{A_r}}{||A_r^T|| ||\widehat{A_r}||} \cdot \frac{B_r^T \widehat{B_r}}{||B_r^T|| ||\widehat{B_r}||}$$

$$w = ||A|| \cdot ||B||$$

Cell Type Abundance Change Simulations

To stimulate experiments where a cell type is not present, we removed cells labeled as a specific cell type and determined if RISE could identify this change. First, RISE was applied to all the cells and conditions of a dataset. Then, the cells of X cell type were removed for a specific experiment, Y, and RISE was applied to this cell type-perturbed dataset. After comparing the canonical genes highly weighted for the gene factors for both analyses, the conditions' component corresponding to canonical markers of the cell type were identified and compared between the two cases. To stimulate an experiment with fewer cells of a certain cell type than other conditions, we subsampled the given cell type by the indicated amount, rather than removing all cells.

Data Normalization, Processing, Cell Annotation, and Visualization

Perturbation Study Analysis

Data Normalization and Processing: The single-cell gene expression was obtained from Chen et al[15]. The first step before normalization was removing doublets with the doubletdetection package applied to the raw RNA counts matrix[98]. Cells with a doublet probability of 0.5 or more were removed from the dataset.

Once doublets were removed, the data was normalized. Genes with less than 0.01 mean raw expression across all cells were removed. Each cell was then normalized by read depth using CPM normalization. Next, each gene was scaled according to its sum. Finally, the data was log transformed via the Delta method,

$$g(y) = \log(y + 1000)$$

where g(y) is the transformed gene and y is the pre-transformed gene[99]. Finally, each gene was mean-centered. The final size of the dataset included 29433 cells and 9461 genes. The data was then fit using the RISE model with 20 components as described above. PaCMAP was used on the concatenated matrix of the projections across all conditions to visualize the cell-to-cell variation and embed cells into low-dimensional space, using default parameters of n_neighbors=10, MN=0.5, and FP_ratio=2.0[50].

Cell Type Annotation: Lower-resolution cellular clusters were defined by Leiden clustering of the RISE projections, resulting in 6 Leiden clusters, with parameters n_neighbors=10 for the nearest neighbors distance and resolution=2.5 for the Leiden algorithm[100]. To provide each cluster with a cell type annotation, genes expressed uniquely within each cluster were found using differential expression analysis by the dispersion-based method with parameters of min_mean=0.005, max_mean=10, min_disp=0.5[101]. These genes were used to annotate each cluster based on their overlap with previously reported marker genes[15]. After the lower-resolution assignments, a finer-grained annotation was performed based on each cluster's expression of a new set of marker genes (Table S5).

SLE Analysis

Data Normalization and Processing: Previous analysis used batch corrected data[62]. To prevent potential sources of bias, we used the raw RNA counts matrix after doublet analysis by the original authors. The same normalization pipeline was used as above, except genes with less than 0.1 mean raw expression across all cells were removed. The final size of the dataset was 1,263,673 cells across all conditions and 2,161 genes. The data was then fit using the RISE model with 30 components as described above.

Cell Type Annotation: Once again, a similar process was used to assign lower and higher resolution cell types as above. Nearest neighbors distance with n_neighbors=10 and resolution=3 for the Leiden algorithm was used to detect 47 Leiden populations. Low and high-resolution cell types were assigned based on a set of marker genes included in Perez et al (Table S5)[62].

Logistic Regression

Scikit learn was used to train an L1-penalized logistic regression model to predict SLE status using the RISE condition factor matrix. A nested cross-validation strategy was initially copied from Perez et al., where the model was trained on the samples of one processing batch and then predicted the rest of the samples in each of the three remaining batches[102]. In addition, an alternative, more traditional nested cross-validation strategy was performed that trained three processing batches and predicted the fourth batch. After the batches are split based on the cross-validation strategy, sample-wise 5-fold inner cross-validation was performed to optimize the L1-penalty of the classifier. Solving was performed with the Stochastic Average Gradient (SAGA) solver, a tolerance of 1×10^{-4}, and a maximum iteration of 10,000[103]. Both the prediction accuracy and ROC AUC were used as a metric to measure the ability to associate the factors with SLE status. Regularization strength was fit by grid search using the cross-validation implementation described above.

Cell Type Composition and Pseudobulking

Cell type composition was analyzed by calculating the percentage of all cells in one experimental condition that corresponded to that cell type. Pseudobulk profiles were created by taking the average of the normalized expression for each cell type in each experimental condition.

Principal Components Analysis

Principal components analysis (PCA) was performed using scikit learn[102]. The data was "flattened" into a two-dimensional matrix of all cells by genes. The number of components was set to match RISE.

Gene Set Enrichment Analysis

The 30 most positively and negatively weighted genes for each component were submitted to ToppGene. ToppGene applies a fuzzy-based similarity measure to calculate semantic similarities between genes. To determine gene annotation significance, p-values were generated using random sampling across the entire genome, based on the empirical distribution of the similarity scores. Significantly (FDR < 0.05) enriched biological processes were retained[104].

REFERENCES

1. Chung, W. *et al.* Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat Commun* **8**, 15081 (2017).
2. Sulen, A., Islam, S., Wolff, A. S. B. & Oftedal, B. E. The prospects of single-cell analysis in autoimmunity. *Scand J Immunol* **92**, (2020).
3. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411–420 (2018).
4. Puram, S. V. *et al.* Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171**, 1611-1624.e24 (2017).
5. Seumois, G. *et al.* Single-cell transcriptomic analysis of allergen-specific T cells in allergy and asthma. *Sci. Immunol.* **5**, eaba6087 (2020).
6. Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).
7. Song, D. & Li, J. J. PseudotimeDE: inference of differential gene expression along cell pseudotime with well-calibrated p-values from single-cell RNA sequencing data. *Genome Biol* **22**, 124 (2021).
8. Yu, L., Cao, Y., Yang, J. Y. H. & Yang, P. Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data. *Genome Biol* **23**, 49 (2022).
9. Ma, Y. *et al.* Integrative differential expression and gene set enrichment analysis using summary statistics for scRNA-seq studies. *Nat Commun* **11**, 1585 (2020).
10. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* **37**, 38–44 (2019).
11. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat Rev Genet* **20**, 257–272 (2019).
12. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* **16**, 1289–1296 (2019).
13. Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**, 1873-1887.e17 (2019).
14. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat Methods* **15**, 1053–1058 (2018).
15. Chen, S. *et al.* Dissecting heterogeneous cell populations across drug and disease conditions with PopAlign. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 28784–28794 (2020).
16. Zhao, J. *et al.* Detection of differentially abundant cell subpopulations in scRNA-seq data. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2100293118 (2021).
17. Chen, W. S. *et al.* Uncovering axes of variation among single-cell cancer specimens. *Nat Methods* **17**, 302–310 (2020).
18. Burkhardt, D. B. *et al.* Quantifying the effect of experimental perturbations at single-cell resolution. *Nat Biotechnol* **39**, 619–629 (2021).
19. Amouzgar, M. *et al.* Supervised dimensionality reduction for exploration of single-cell data by HSS-LDA. *Patterns* **3**, 100536 (2022).
20. Loos, C., Moeller, K., Fröhlich, F., Hucho, T. & Hasenauer, J. A Hierarchical, Data-Driven Approach to Modeling Single-Cell Populations Predicts Latent Causes of Cell-To-Cell Variability. *Cell Systems* **6**, 593-603.e13 (2018).
21. Barkley, D. *et al.* Cancer cell states recur across tumor types and form specific interactions with the tumor microenvironment. *Nat Genet* **54**, 1192–1201 (2022).

22. Tan, Z. C. & Meyer, A. S. The structure is the message: preserving experimental context through tensor decomposition.

23. Kolda, T. G. & Bader, B. W. Tensor Decompositions and Applications. *SIAM Review* **51**, 455–500 (2009).

24. Tan, Z. C., Murphy, M. C., Alpay, H. S., Taylor, S. D. & Meyer, A. S. Tensor-structured decomposition improves systems serology analysis. *Mol Syst Biol* **17**, (2021).

25. Hore, V. *et al.* Tensor decomposition for multiple-tissue gene expression experiments. *Nat Genet* **48**, 1094–1100 (2016).

26. Orcutt-Jahns, B., Emmel, P. C., Snyder, E. M., Taylor, S. D. & Meyer, A. S. *Multivalent, Asymmetric IL-2-Fc Fusions Provide Optimally Enhanced Regulatory T Cell Selectivity*. http://biorxiv.org/lookup/doi/10.1101/2021.07.03.451002 (2021) doi:10.1101/2021.07.03.451002.

27. Chin, J. L. *et al.* Cytokine-expression patterns reveal coordinated immunological programs associated with persistent MRSA bacteremia. Preprint at https://doi.org/10.1101/2022.12.28.521386 (2022).

28. Armingol, E. *et al.* Context-aware deconvolution of cell–cell communication with Tensor-cell2cell. *Nat Commun* **13**, 3665 (2022).

29. Martino, C. *et al.* Context-aware dimensionality reduction deconvolutes gut microbial community dynamics. *Nat Biotechnol* **39**, 165–168 (2021).

30. Orcutt-Jahns, B. *et al. Multivalency Enhances the Specificity of Fc-Cytokine Fusions*. http://biorxiv.org/lookup/doi/10.1101/2021.07.03.451002 (2021) doi:10.1101/2021.07.03.451002.

31. Farhat, A. M. *et al.* Modeling cell-specific dynamics and regulation of the common gamma chain cytokines. *Cell Reports* **35**, 109044 (2021).

32. Acar, E., Kolda, T. G. & Dunlavy, D. M. All-at-once Optimization for Coupled Matrix and Tensor Factorizations. *arXiv:1105.3422 [physics, stat]* (2011).

33. Acar, E., Bro, R. & Smilde, A. K. Data Fusion in Metabolomics Using Coupled Matrix and Tensor Factorizations. *Proceedings of the IEEE* **103**, 1602–1620 (2015).

34. Borsoi, R. A. *et al.* Coupled CP tensor decomposition with shared and distinct components for multi-task fMRI data fusion. Preprint at http://arxiv.org/abs/2211.14253 (2022).

35. Chin, J. L., Chan, L. C., Yeaman, M. R. & Meyer, A. S. Tensor-based insights into systems immunity and infectious disease. *Trends in Immunology* **44**, 329–332 (2023).

36. Chin, J. L. *et al. Cytokine-Expression Patterns Reveal Coordinated Immunological Programs Associated with Persistent MRSA Bacteremia*. http://biorxiv.org/lookup/doi/10.1101/2022.12.28.521386 (2022) doi:10.1101/2022.12.28.521386.

37. Kiers, H. A. L., ten Berge, J. M. F. & Bro, R. PARAFAC2—Part I. A direct fitting algorithm for the PARAFAC2 model. *J. Chemometrics* **13**, 275–294 (1999).

38. Augustijn, D., Kulakova, A., Mahapatra, S., Harris, P. & Rinnan, Å. Isothermal Chemical Denaturation: Data Analysis, Error Detection, and Correction by PARAFAC2. *Anal. Chem.* **92**, 6958–6967 (2020).

39. Amante *et al.* Untargeted Metabolomic Profile for the Detection of Prostate Carcinoma—Preliminary Results from PARAFAC2 and PLS–DA Models. *Molecules* **24**, 3063 (2019).

40. Ren, Y., Lou, J., Xiong, L. & Ho, J. C. Robust Irregular Tensor Factorization and Completion for Temporal Health Data Analysis. in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* 1295–1304 (ACM, Virtual Event Ireland, 2020). doi:10.1145/3340531.3411982.

41. Zhang, R., Zhou, T. & Ma, J. Ultrafast and interpretable single-cell 3D genome analysis with Fast-Higashi. *Cell Systems* **13**, 798-807.e6 (2022).

42. Yahyanejad, F., Albert, R. & DasGupta, B. A survey of some tensor analysis techniques for biological systems. *Quant. Biol.* **7**, 266–277 (2019).

43. Bro, R., Andersson, C. A. & Kiers, H. A. L. PARAFAC2—Part II. Modeling chromatographic data with retention time shifts. *J. Chemometrics* **13**, 295–309 (1999).

44. Cattell, R. B. 'PARAI,I,EL PROPORTIONAL PROFILES' AND OTHER PRINCIPLES FOR DETERMINING THE CHOICE OF FACTORSB.YROTATION.

45. Kroonenberg, P. M. & Harshman, R. A. CATTELL'S parallel proportional profiles: The triumph of a prodigal rotation. *Journal of Chemometrics* **35**, e3235 (2021).

46. Bro, R. & Kiers, H. A. L. A new efficient method for determining the number of components in PARAFAC models. *Journal of Chemometrics* **17**, 274–286 (2003).

47. Tian, K., Wu, L., Min, S. & Bro, R. Geometric search: A new approach for fitting PARAFAC2 models on GC-MS data. *Talanta* **185**, 378–386 (2018).

48. Helwig, N. E. Estimating latent trends in multivariate longitudinal data via Parafac2 with functional and structural constraints: Parafac2 with functional and structural constraints. *Biom. J.* **59**, 783–803 (2017).

49. Roald, M. *et al.* An AO-ADMM Approach to Constraining PARAFAC2 on All Modes. *SIAM Journal on Mathematics of Data Science* **4**, 1191–1222 (2022).

50. Wang, Y., Huang, H., Rudin, C. & Shaposhnik, Y. Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap, and PaCMAP for Data Visualization.

51. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* **19**, 41–50 (2022).

52. Thul, P. J. *et al.* A subcellular map of the human proteome. *Science* **356**, eaal3321 (2017).

53. Ghanem, M. H. *et al.* Proteomic and Single-Cell Transcriptomic Dissection of Human Plasmacytoid Dendritic Cell Response to Influenza Virus. *Front Immunol* **13**, 814627 (2022).

54. Vescovi, R. *et al.* Collapse of the Plasmacytoid Dendritic Cell Compartment in Advanced Cutaneous Melanomas by Components of the Tumor Cell Secretome. *Cancer Immunology Research* **7**, 12–28 (2019).

55. Inoue, H., Aihara, M., Tomioka, M. & Watabe, Y. Specific Enhancement of Vascular Endothelial Growth Factor (VEGF) Production in Ischemic Region by Alprostadil — Potential Therapeutic Application in Pharmaceutical Regenerative Medicine. *Journal of Pharmacological Sciences* **122**, 158–161 (2013).

56. Haider, D. G. *et al.* PGE1 analog alprostadil induces VEGF and eNOS expression in endothelial cells. *American Journal of Physiology-Heart and Circulatory Physiology* **289**, H2066–H2072 (2005).

57. Oshima, H. *et al.* Activation of epidermal growth factor receptor signaling by the prostaglandin E2 receptor EP4 pathway during gastric tumorigenesis. *Cancer Science* **102**, 713–719 (2011).

58. Zhang, K. *et al.* Endogenous glucocorticoids promote the expansion of myeloid-derived suppressor cells in a murine model of trauma. *Int J Mol Med* **30**, 277–282 (2012).

59. Varga, G. *et al.* Glucocorticoids induce an activated, anti-inflammatory monocyte subset in mice that resembles myeloid-derived suppressor cells. *Journal of Leukocyte Biology* **84**, 644–650 (2008).

60. Lu, Y. *et al.* Glucocorticoid receptor promotes the function of myeloid-derived suppressor cells by suppressing HIF1α-dependent glycolysis. *Cell Mol Immunol* **15**, 618–629 (2018).

61. Silva-Gomes, R. *et al.* Differential expression and regulation of MS4A family members in myeloid cells in physiological and pathological conditions. *J Leukoc Biol* **111**, 817–836 (2022).

62. Perez, R. K. *et al.* Single-cell RNA-seq reveals cell type–specific molecular and genetic associations to lupus. *Science* **376**, eabf1970 (2022).

63. Steinke, F. C. *et al.* TCF-1 and LEF-1 act upstream of Th-POK to promote the CD4+ T cell fate and interact with Runx3 to silence Cd4 in CD8+ T cells. *Nat Immunol* **15**, 646–656 (2014).

64. Hwang, J.-R., Byeon, Y., Kim, D. & Park, S.-G. Recent insights of T cell receptor-mediated signaling pathways for T cell activation and development. *Exp Mol Med* **52**, 750–761 (2020).

65. Zheng, M. *et al.* Single-cell sequencing shows cellular heterogeneity of cutaneous lesions in lupus erythematosus. *Nat Commun* **13**, 7489 (2022).

66. Xiong, H. *et al.* Cytotoxic CD161−CD8+ TEMRA cells contribute to the pathogenesis of systemic lupus erythematosus. *eBioMedicine* **90**, 104507 (2023).

67. Li, H., Boulougoura, A., Endo, Y. & Tsokos, G. C. Abnormalities of T cells in systemic lupus erythematosus: new insights in pathogenesis and therapeutic strategies. *Journal of Autoimmunity* **132**, 102870 (2022).

68. Kim, J.-W. *et al.* S100A8 in Serum, Urine, and Saliva as a Potential Biomarker for Systemic Lupus Erythematosus. *Front. Immunol.* **13**, 886209 (2022).

69. Kitagori, K. *et al.* Expression of S100A8 protein on B cells is associated with disease activity in patients with systemic lupus erythematosus. *Arthritis Res Ther* **25**, 76 (2023).

70. Lood, C. *et al.* Protein synthesis of the pro-inflammatory S100A8/A9 complex in plasmacytoid dendritic cells and cell surface S100A8/A9 on leukocyte subpopulations in systemic lupus erythematosus. *Arthritis Res Ther* **13**, R60 (2011).

71. Muñoz-Grajales, C. *et al.* Serum S100A8/A9 and MMP-9 levels are elevated in systemic lupus erythematosus patients with cognitive impairment. *Front. Immunol.* **14**, 1326751 (2024).

72. Demers-Mathieu, V. Optimal Selection of IFN-α-Inducible Genes to Determine Type I Interferon Signature Improves the Diagnosis of Systemic Lupus Erythematosus. *Biomedicines* **11**, 864 (2023).

73. Jin, Z. *et al.* Single-cell gene expression patterns in lupus monocytes independently indicate disease activity, interferon and therapy. *Lupus Sci Med* **4**, e000202 (2017).

74. Zhao, X. *et al.* Identification of key biomarkers and immune infiltration in systemic lupus erythematosus by integrated bioinformatics analysis. *J Transl Med* **19**, 35 (2021).

75. Lotfollahi, M. *et al.* Mapping single-cell data to reference atlases by transfer learning. *Nat Biotechnol* **40**, 121–130 (2022).

76. Xu, C. *et al.* Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular Systems Biology* **17**, e9620 (2021).

77. Vermi, W. *et al.* Role of ChemR23 in directing the migration of myeloid and plasmacytoid dendritic cells to lymphoid organs and inflamed skin. *The Journal of Experimental Medicine* **201**, 509–515 (2005).

78. Lu, Y. *et al.* Glucocorticoid receptor promotes the function of myeloid-derived suppressor cells by suppressing HIF1α-dependent glycolysis. *Cell Mol Immunol* **15**, 618–629 (2018).

79. Hirose, S., Lin, Q., Ohtsuji, M., Nishimura, H. & Verbeek, J. S. Monocyte subsets involved in the development of systemic lupus erythematosus and rheumatoid arthritis. *International Immunology* **31**, 687–696 (2019).

80. Paredes, J. L., Fernandez-Ruiz, R. & Niewold, T. B. T Cells in Systemic Lupus Erythematosus. *Rheumatic Disease Clinics of North America* **47**, 379–393 (2021).

81. Chen, P.-M. & Tsokos, G. C. The role of CD8+ T-cell systemic lupus erythematosus pathogenesis: an update. *Current Opinion in Rheumatology* **33**, 586–591 (2021).

82. Lu, Z. *et al.* Increased oxidative stress contributes to impaired peripheral CD56dimCD57+ NK cells from patients with systemic lupus erythematosus. *Arthritis Res Ther* **24**, 48 (2022).

83. Bing, P.-F. *et al.* Common Marker Genes Identified from Various Sample Types for Systemic Lupus Erythematosus. *PLoS ONE* **11**, e0156234 (2016).

84. Zhong, Y. *et al.* Screening Biomarkers for Systemic Lupus Erythematosus Based on Machine Learning and Exploring Their Expression Correlations With the Ratios of Various Immune Cells. *Front. Immunol.* **13**, 873787 (2022).

85. Siddiqi, K. Z., Wilhelm, T. R., Ulff-Møller, C. J. & Jacobsen, S. Cluster of highly expressed interferon-stimulated genes associate more with African ancestry than disease activity in patients with systemic lupus erythematosus. A systematic review of cross-sectional studies. *Translational Research* **238**, 63–75 (2021).

86. Eggenhuizen, P. J. *et al.* Smith-specific regulatory T cells halt the progression of lupus nephritis. *Nat Commun* **15**, 899 (2024).

87. Geuenich, M. J., Gong, D. & Campbell, K. R. The impacts of active and self-supervised learning on efficient annotation of single-cell expression data. *Nat Commun* **15**, 1014 (2024).

88. Zhang, Z. *et al.* Signal recovery in single cell batch integration. Preprint at https://doi.org/10.1101/2023.05.05.539614 (2023).

89. Owen, A. B. & Perry, P. O. Bi-cross-validation of the SVD and the nonnegative matrix factorization. *Ann. Appl. Stat.* **3**, (2009).

90. Cohen, J. E. & Bro, R. Nonnegative PARAFAC2: a flexible coupling approach. Preprint at http://arxiv.org/abs/1802.05035 (2018).

91. Ren, Y. *et al.* MULTIPAR: Supervised Irregular Tensor Factorization with Multi-task Learning. Preprint at http://arxiv.org/abs/2208.00993 (2022).

92. Sorochan Armstrong, M. D., Hinrich, J. L., De La Mata, A. P. & Harynuk, J. J. PARAFAC2×N: Coupled decomposition of multi-modal data with drift in N modes. *Analytica Chimica Acta* **1249**, 340909 (2023).

93. Fang, S. *et al.* Computational Approaches and Challenges in Spatial Transcriptomics. *Genomics, Proteomics & Bioinformatics* **21**, 24–47 (2023).

94. Bro, R. PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems* **38**, 149–171 (1997).

95. Halko, N., Martinsson, P.-G. & Tropp, J. A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. Preprint at http://arxiv.org/abs/0909.4061 (2010).

96. Kossaifi, J., Panagakis, Y., Anandkumar, A. & Pantic, M. TensorLy: Tensor Learning in Python. *Journal of Machine Learning Research* **20**, 1–6 (2019).

97. Yu, H., Augustijn, D. & Bro, R. Accelerating PARAFAC2 algorithms for non-negative complex tensor decomposition. *Chemometrics and Intelligent Laboratory Systems* **214**, 104312 (2021).

98. Xi, N. M. & Li, J. J. Protocol for executing and benchmarking eight computational doublet-detection methods in single-cell RNA sequencing data analysis. *STAR Protocols* **2**, 100699 (2021).

99. Ahlmann-Eltze, C. & Huber, W. Comparison of transformations for single-cell RNA-seq data. *Nat Methods* (2023) doi:10.1038/s41592-023-01814-1.

100. Traag, V. A., Waltman, L. & Van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* **9**, 5233 (2019).

101. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33**, 495–502 (2015).

102. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*.

103. Defazio, A., Bach, F. & Lacoste-Julien, S. SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives.

104. Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Research* **37**, W305–W311 (2009).

Figure 1

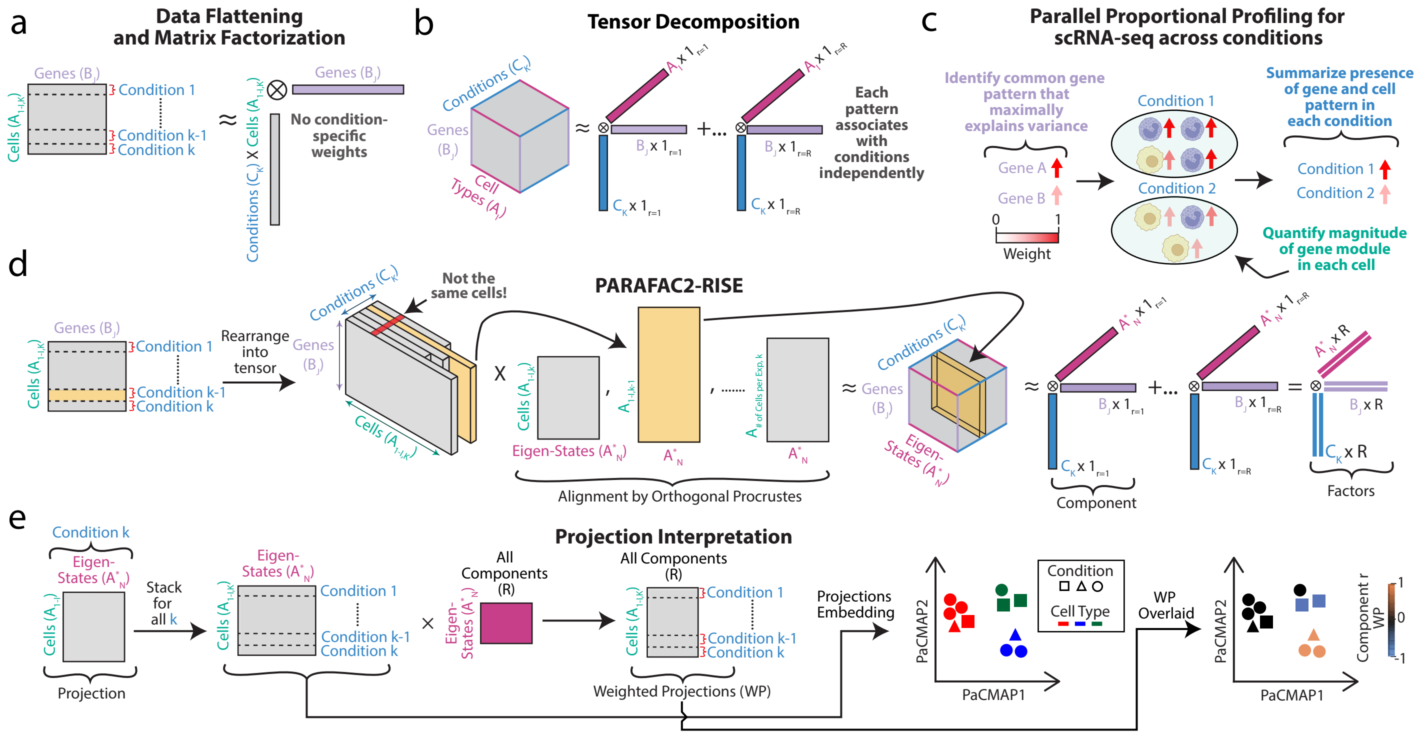

a
Data Flattening and Matrix Factorization
Genes (B_J)
Cells (A_1-I,K)
Condition 1
Condition k-1
Condition k
≈
Cells (A_1-I,K)
Conditions (C_K) X Cells (A_1-I,K)
Genes (B_J)
No condition-specific weights
b
Tensor Decomposition
Conditions (C_K)
Genes (B_J)
Cell Types (A_I)
≈
A_I x 1 r=1
B_J x 1 r=1
C_K x 1 r=1
+...
A_I x 1 r=R
B_J x 1 r=R
C_K x 1 r=R
Each pattern associates with conditions independently
c
Parallel Proportional Profiling for scRNA-seq across conditions
Identify common gene pattern that maximally explains variance
Gene A
Gene B
0
1
Weight
Condition 1
Condition 2
Summarize presence of gene and cell pattern in each condition
Condition 1
Condition 2
Quantify magnitude of gene module in each cell
d
Genes (B_J)
Cells (A_1-I,K)
Condition 1
Condition k-1
Condition k
Rearrange into tensor
Conditions (C_K)
Genes (B_J)
Cells (A_1-I,K)
Not the same cells!
PARAFAC2-RISE
X
Cells (A_1-I,K)
Eigen-States (A*_N)
A_1-I,k-1
A*_N
A_# of Cells per Exp. k
A*_N
Alignment by Orthogonal Procrustes
≈
Conditions (C_K)
Genes (B_J)
Eigen-States (A*_N)
≈
A*_N x 1 r=1
B_J x 1 r=1
C_K x 1 r=1
Component
+...
A*_N x 1 r=R
B_J x 1 r=R
C_K x 1 r=R
=
A*_N x R
B_J x R
C_K x R
Factors
e
Projection Interpretation
Condition k
Eigen-States (A*_N)
Cells (A_1-I)
Projection
Stack for all k
Eigen-States (A*_N)
Cells (A_1-I,K)
Condition 1
Condition k-1
Condition k
×
All Components (R)
Eigen-States (A*_N)
All Components (R)
Cells (A_1-I,K)
Condition 1
Condition k-1
Condition k
Weighted Projections (WP)
Projections Embedding
PaCMAP2
PaCMAP1
Condition
Cell Type
WP Overlaid
PaCMAP2
PaCMAP1
Component r WP
1
0
-1

Figure 2

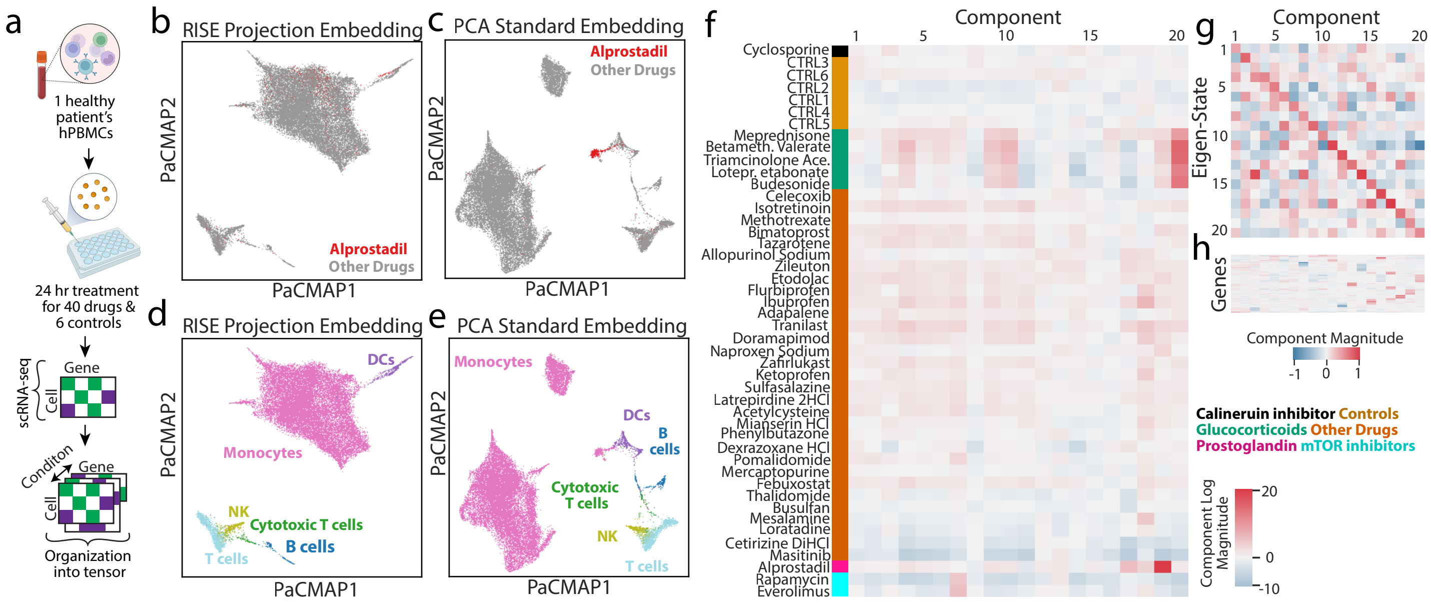
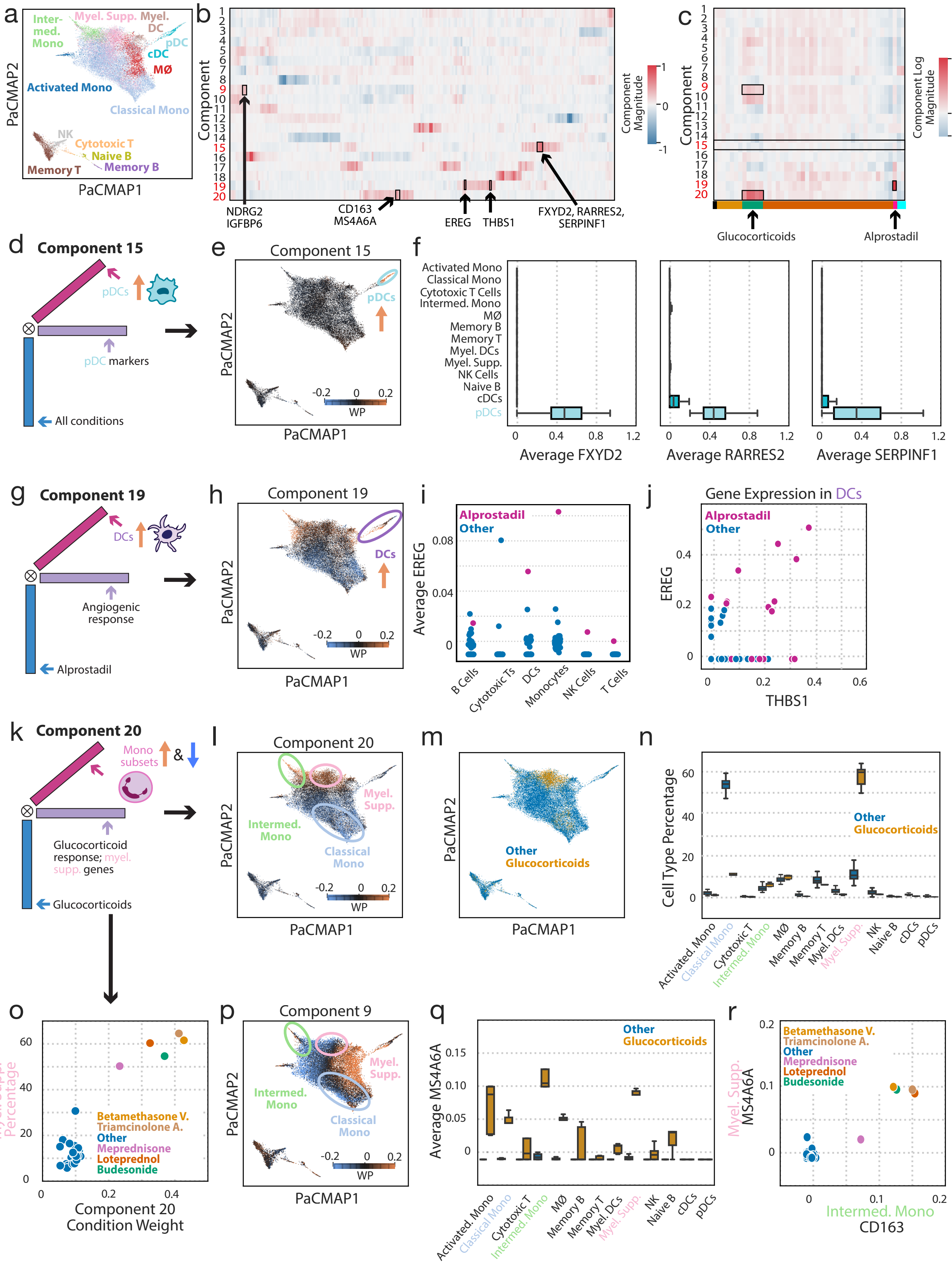

a
1 healthy patient's hPBMCs
24 hr treatment for 40 drugs & 6 controls
scRNA-seq
Gene
Cell
Conditon
Gene
Cell
Organization into tensor
b
RISE Projection Embedding
PaCMAP2
PaCMAP1
Alprostadil
Other Drugs
c
PCA Standard Embedding
PaCMAP2
PaCMAP1
Alprostadil
Other Drugs
d
RISE Projection Embedding
PaCMAP2
PaCMAP1
DCs
Monocytes
NK
Cytotoxic T cells
B cells
T cells
e
PCA Standard Embedding
PaCMAP2
PaCMAP1
Monocytes
DCs
B cells
Cytotoxic T cells
NK
T cells
f
Component
1
5
10
15
20
Cyclosporine
CTRL3
CTRL6
CTRL2
CTRL1
CTRL4
CTRL5
Meprednisone
Betameth. Valerate
Triamcinolone Ace.
Lotepr. etabonate
Budesonide
Celecoxib
Isotretinoin
Methotrexate
Bimatoprost
Tazarotene
Allopurinol Sodium
Zileuton
Etodolac
Flurbiprofen
Ibuprofen
Adapalene
Tranilast
Doramapimod
Naproxen Sodium
Zafirlukast
Ketoprofen
Sulfasalazine
Latrepirdine 2HCl
Acetylcysteine
Mianserin HCl
Phenylbutazone
Dexrazoxane HCl
Pomalidomide
Mercaptopurine
Febuxostat
Thalidomide
Busulfan
Mesalamine
Loratadine
Cetirizine DiHCl
Masitinib
Alprostadil
Rapamycin
Everolimus
g
Component
1
5
10
15
20
Eigen-State
1
5
10
15
20
h
Genes
Component Magnitude
-1
0
1
Calineruin inhibitor
Controls
Glucocorticoids
Other Drugs
Prostoglandin
mTOR inhibitors
Component Log Magnitude
20
0
-10

Figure 3



a
Inter-med. Mono
Myel. Supp.
Myel. DC
pDC
cDC
MØ
Activated Mono
Classical Mono
NK
Cytotoxic T
Naive B
Memory T
Memory B
PaCMAP2
PaCMAP1
b
Component
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
NDRG2
IGFBP6
CD163
MS4A6A
EREG
THBS1
FXYD2, RARRES2, SERPINF1
Component Magnitude
1
0
-1
c
Component
Component Log Magnitude
20
0
-10
Glucocorticoids
Alprostadil
d
Component 15
pDCs
pDC markers
All conditions
e
Component 15
pDCs
PaCMAP2
PaCMAP1
-0.2
0
0.2
WP
f
Activated Mono
Classical Mono
Cytotoxic T Cells
Intermed. Mono
MØ
Memory B
Memory T
Myel. DCs
Myel. Supp.
NK Cells
Naive B
cDCs
pDCs
0
0.4
0.8
1.2
Average FXYD2
Average RARRES2
Average SERPINF1
g
Component 19
DCs
Angiogenic response
Alprostadil
h
Component 19
DCs
PaCMAP2
PaCMAP1
-0.2
0
0.2
WP
i
Alprostadil
Other
Average EREG
0
0.04
0.08
B Cells
Cytotoxic Ts
DCs
Monocytes
NK Cells
T Cells
j
Gene Expression in DCs
Alprostadil
Other
EREG
0
0.2
0.4
0
0.2
0.4
0.6
THBS1
k
Component 20
Mono subsets
&
Glucocorticoid response; myel. supp. genes
Glucocorticoids
l
Component 20
Intermed. Mono
Myel. Supp.
Classical Mono
PaCMAP2
PaCMAP1
-0.2
0
0.2
WP
m
Other
Glucocorticoids
PaCMAP2
PaCMAP1
n
Cell Type Percentage
0
20
40
60
Other
Glucocorticoids
Activated. Mono
Classical Mono
Cytotoxic T
Intermed. Mono
MØ
Memory B
Memory T
Myel. DCs
Myel. Supp.
NK
Naive B
cDCs
pDCs
o
Myel. Supp. Percentage
0
20
40
60
0
0.2
0.4
Component 20 Condition Weight
Betamethasone V.
Triamcinolone A.
Other
Meprednisone
Loteprednol
Budesonide
p
Component 9
Intermed. Mono
Myel. Supp.
Classical Mono
PaCMAP2
PaCMAP1
-0.2
0
0.2
WP
q
Other
Glucocorticoids
Average MS4A6A
0
0.05
0.10
0.15
Activated. Mono
Classical Mono
Cytotoxic T
Intermed. Mono
MØ
Memory B
Memory T
Myel. DCs
Myel. Supp.
NK
Naive B
cDCs
pDCs
r
Betamethasone V.
Triamcinolone A.
Other
Meprednisone
Loteprednol
Budesonide
Myel. Supp. MS4A6A
0
0.1
0.2
0
0.1
0.2
Intermed. Mono CD163

Figure 4

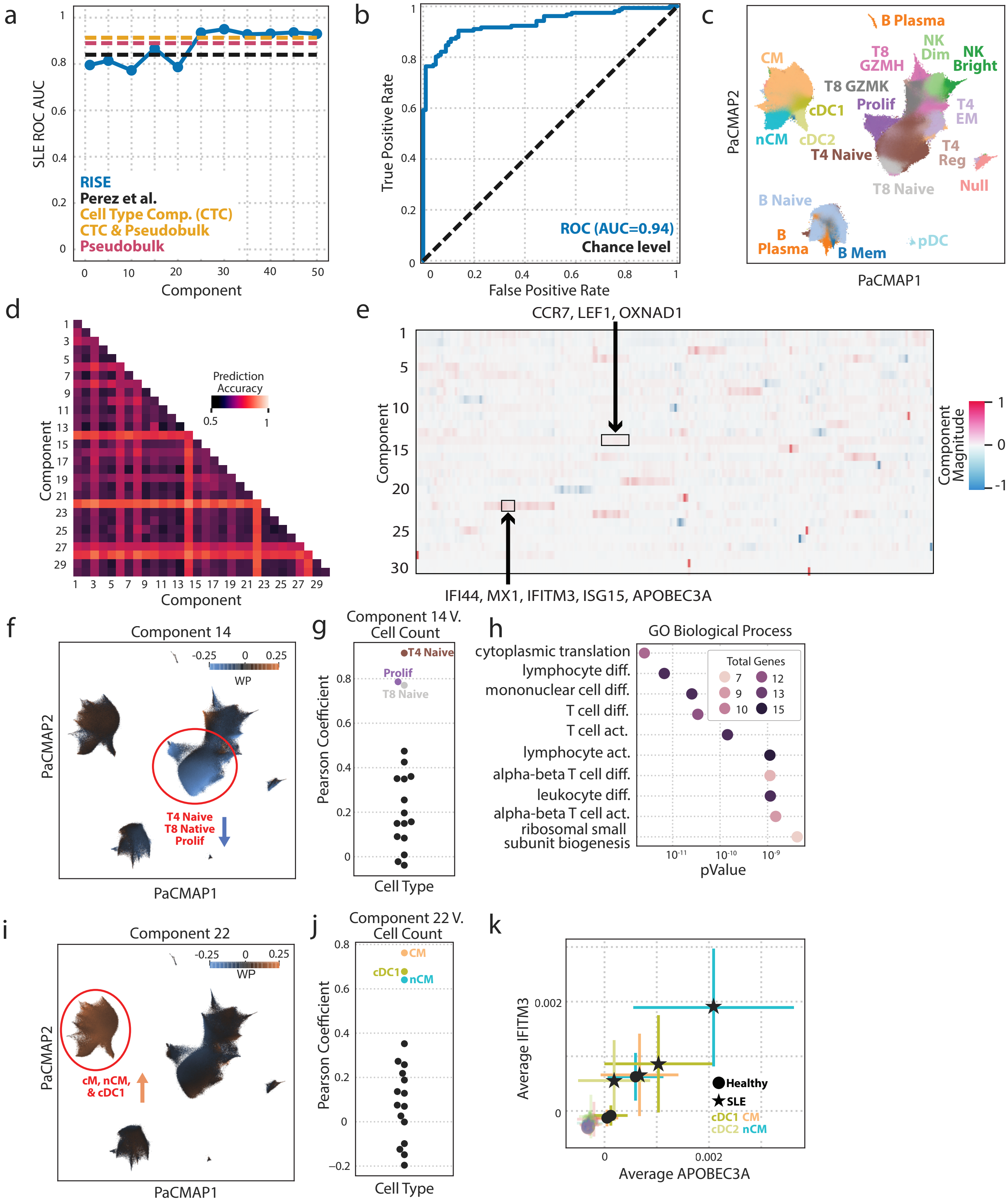

a
SLE ROC AUC
1
0.8
0.6
0.4
0.2
0
RISE
Perez et al.
Cell Type Comp. (CTC)
CTC & Pseudobulk
Pseudobulk
0
10
20
30
40
50
Component
b
True Positive Rate
1
0.8
0.6
0.4
0.2
0
ROC (AUC=0.94)
Chance level
0
0.2
0.4
0.6
0.8
1
False Positive Rate
c
B Plasma
CM
T8 GZMH
NK Dim
NK Bright
T8 GZMK
cDC1
Prolif
T4 EM
nCM
cDC2
T4 Naive
T4 Reg
T8 Naive
Null
B Naive
B Plasma
B Mem
pDC
PaCMAP2
PaCMAP1
d
Component
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
Prediction Accuracy
0.5
1
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
Component
e
CCR7, LEF1, OXNAD1
Component
1
5
10
15
20
25
30
1
0
-1
Component Magnitude
IFI44, MX1, IFITM3, ISG15, APOBEC3A
f
Component 14
-0.25
0
0.25
WP
PaCMAP2
T4 Naive
T8 Native
Prolif
PaCMAP1
g
Component 14 V. Cell Count
T4 Naive
Prolif
T8 Naive
Pearson Coefficient
0.8
0.6
0.4
0.2
0
Cell Type
h
GO Biological Process
cytoplasmic translation
lymphocyte diff.
mononuclear cell diff.
T cell diff.
T cell act.
lymphocyte act.
alpha-beta T cell diff.
leukocyte diff.
alpha-beta T cell act.
ribosomal small subunit biogenesis
Total Genes
7
9
10
12
13
15
10^{-11}
10^{-10}
10^{-9}
pValue
i
Component 22
-0.25
0
0.25
WP
PaCMAP2
cM, nCM, & cDC1
PaCMAP1
j
Component 22 V. Cell Count
CM
cDC1
nCM
Pearson Coefficient
0.8
0.6
0.4
0.2
0
−0.2
Cell Type
k
Average IFITM3
0.002
0
Healthy
SLE
cDC1
CM
cDC2
nCM
0
0.002
Average APOBEC3A

Figure 5

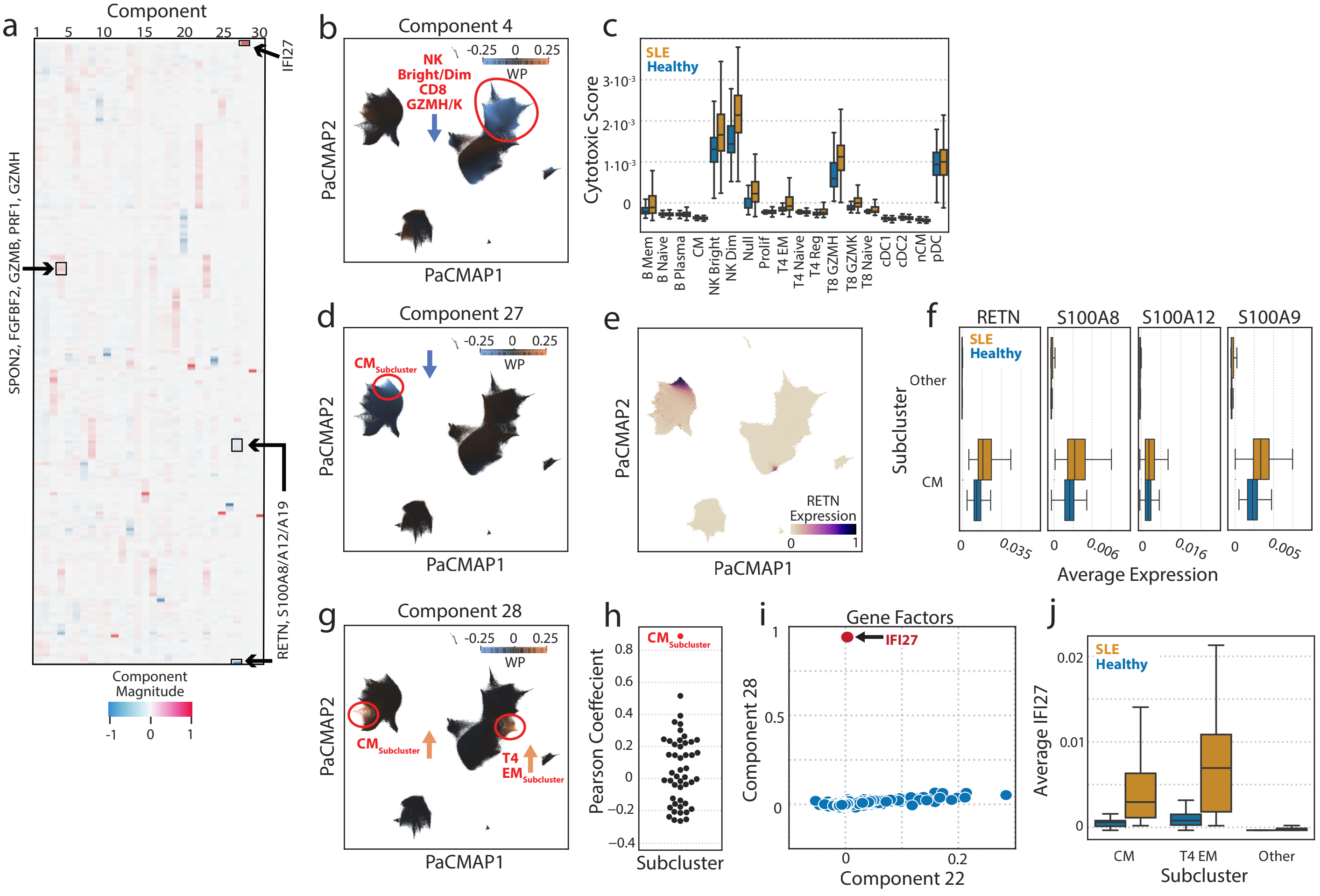
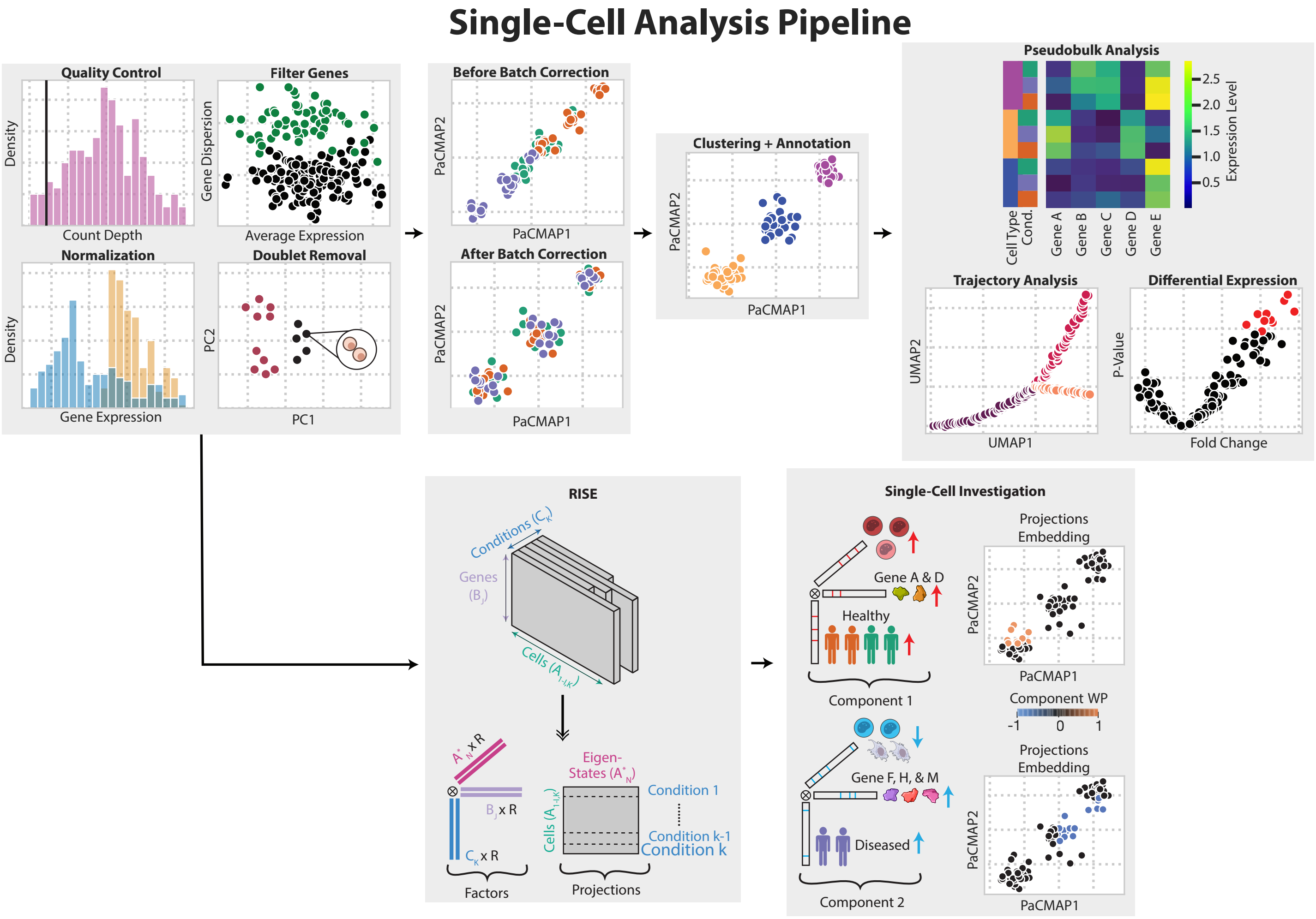

a
Component
1 5 10 15 20 25 30
IFI27
SPON2, FGFBF2, GZMB, PRF1, GZMH
RETN, S100A8/A12/A19
Component Magnitude
-1 0 1
b
Component 4
-0.25 0 0.25
WP
NK
Bright/Dim
CD8
GZMH/K
PaCMAP2
PaCMAP1
c
SLE
Healthy
Cytotoxic Score
3·10⁻³
2·10⁻³
1·10⁻³
0
B Mem
B Naive
B Plasma
CM
NK Bright
NK Dim
Null
Prolif
T4 EM
T4 Naive
T4 Reg
T8 GZMH
T8 GZMK
T8 Naive
cDC1
cDC2
nCM
pDC
d
Component 27
-0.25 0 0.25
WP
CM Subcluster
PaCMAP2
PaCMAP1
e
PaCMAP2
PaCMAP1
RETN
Expression
0 1
f
RETN
S100A8
S100A12
S100A9
SLE
Healthy
Subcluster
Other
CM
0 0.035
0 0.006
0 0.016
0 0.005
Average Expression
g
Component 28
-0.25 0 0.25
WP
CM Subcluster
T4
EM Subcluster
PaCMAP2
PaCMAP1
h
CM Subcluster
Pearson Coeffecient
0.8
0.6
0.4
0.2
0
−0.2
−0.4
Subcluster
i
Gene Factors
IFI27
Component 28
1
0.5
0
0
0.2
Component 22
j
SLE
Healthy
Average IFI27
0.02
0.01
0
CM
T4 EM
Other
Subcluster

Figure 6


Single-Cell Analysis Pipeline
Quality Control
Density
Count Depth
Filter Genes
Gene Dispersion
Average Expression
Normalization
Density
Gene Expression
Doublet Removal
PC2
PC1
Before Batch Correction
PaCMAP2
PaCMAP1
After Batch Correction
PaCMAP2
PaCMAP1
Clustering + Annotation
PaCMAP2
PaCMAP1
Pseudobulk Analysis
Cell Type
Cond.
Gene A
Gene B
Gene C
Gene D
Gene E
2.5
2.0
1.5
1.0
0.5
Expression Level
Trajectory Analysis
UMAP2
UMAP1
Differential Expression
P-Value
Fold Change
RISE
Conditions (C_K)
Genes (B_J)
Cells (A_1-I,K)
A*_N x R
B_J x R
C_K x R
Factors
Eigen-States (A*_N)
Cells (A_1-I,K)
Condition 1
Condition k-1
Condition k
Projections
Single-Cell Investigation
Gene A & D
Healthy
Component 1
Projections Embedding
PaCMAP2
PaCMAP1
Component WP
-1
0
1
Gene F, H, & M
Diseased
Component 2
Projections Embedding
PaCMAP2
PaCMAP1