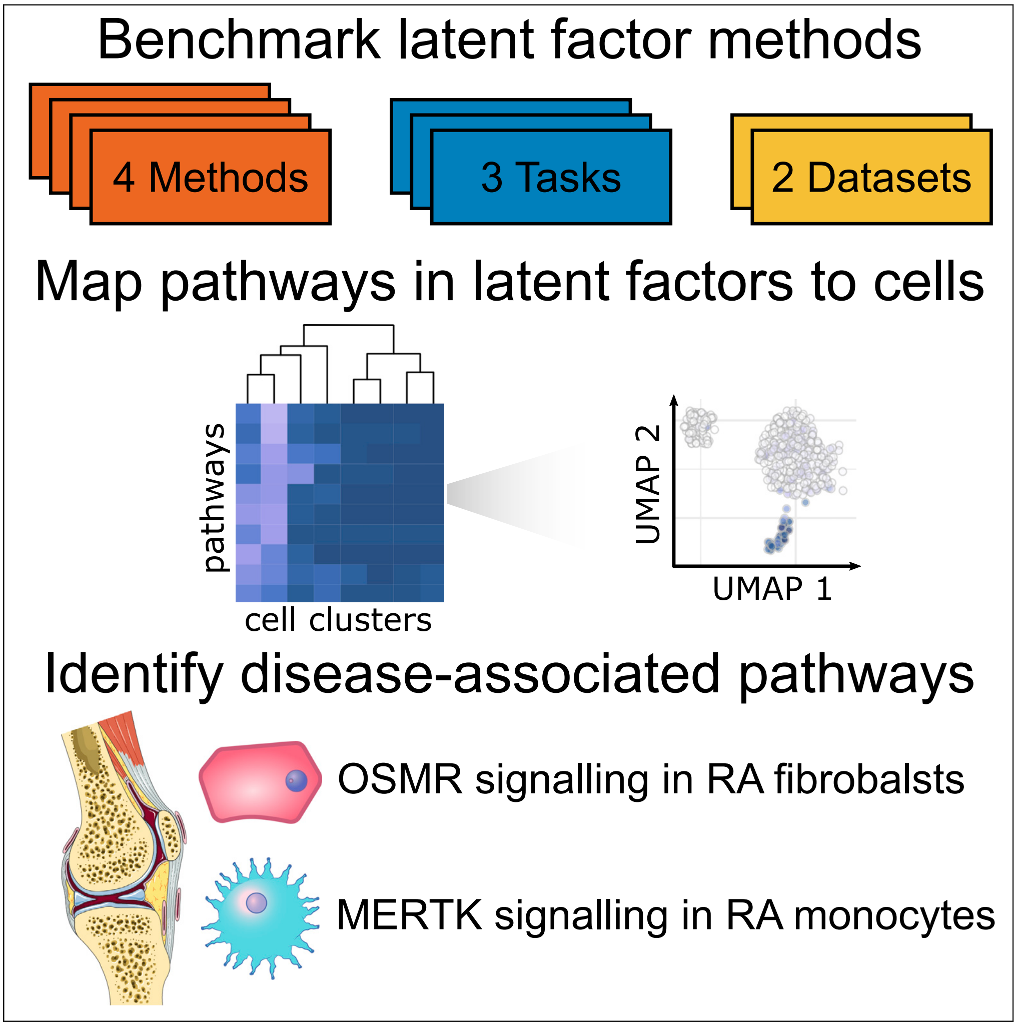
iScience

CellPress
OPEN ACCESS

Article

Latent Factor Modeling of scRNA-Seq Data Uncovers Dysregulated Pathways in Autoimmune Disease Patients


Benchmark latent factor methods
4 Methods
3 Tasks
2 Datasets
Map pathways in latent factors to cells
pathways
cell clusters
UMAP 2
UMAP 1
Identify disease-associated pathways
OSMR signalling in RA fibrobalsts
MERTK signalling in RA monocytes

Giovanni Palla,
Enrico Ferrero

enrico.ferrero@novartis.com

HIGHLIGHTS

Benchmarking four latent factor models for analysis of scRNA-seq data

Map biological pathways associated with latent factors to specific cell subsets

OSMR pathway dysregulated in a subset of RA synovial fibroblasts

MERTK-expressing monocytes with efferocytic function depleted in RA

Palla & Ferrero, iScience *23*, 101451
September 25, 2020 © 2020 The Authors.
https://doi.org/10.1016/j.isci.2020.101451

iScience

CellPress OPEN ACCESS

Article

Latent Factor Modeling of scRNA-Seq Data Uncovers Dysregulated Pathways in Autoimmune Disease Patients

Giovanni Palla[1,2] and Enrico Ferrero[1,3,*]

SUMMARY

Latent factor modeling applied to single-cell RNA sequencing (scRNA-seq) data is a useful approach to discover gene signatures. However, it is often unclear what methods are best suited for specific tasks and how latent factors should be interpreted.

Here, we compare four state-of-the-art methods and propose an approach to assign derived latent factors to pathway activities and specific cell subsets. By applying this framework to scRNA-seq datasets from biopsies of patients with rheumatoid arthritis and systemic lupus erythematosus, we discover disease-relevant gene signatures in specific cellular subsets. In rheumatoid arthritis, we identify an inflammatory OSMR signaling signature active in a subset of synovial fibroblasts and an efferocytic signature in a subset of synovial monocytes.

Overall, we provide insights into latent factors models for the analysis of scRNA-seq data, develop a framework to identify cell subtypes in a phenotype-driven way, and use it to identify novel pathways dysregulated in rheumatoid arthritis.

INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) is a powerful technique that enables gene expression measurements in thousands of individual cells. Resolving cellular heterogeneity by scRNA-seq has enabled groundbreaking discovery in the biomedical domain, such as finding key disease drivers in cancer (Patel et al., 2014; Puram et al., 2017; Tirosh et al., 2016), neurodegeneration (Keren-Shaul et al., 2017; Mathys et al., 2019), and immune-mediated diseases (Der et al., 2019; Smillie et al., 2019; The Accelerating Medicines Partnership in SLE Network et al., 2019; Zhang et al., 2019). From a data analysis standpoint, a crucial step in a standard scRNA-seq pipeline is clustering (Luecken and Theis, 2019), where discrete cell populations sharing a common transcriptional profile are defined. These cell clusters are used in a variety of downstream analyses, such as differential expression (Crowell et al., 2019; Ma et al., 2019; Soneson and Robinson, 2018), compositional analysis (Fonseka et al., 2018), and cellular interaction analysis (Vento-Tormo et al., 2018; Yuan et al., 2019; Zhou et al., 2017). Phenotypic identification of clusters is usually performed by means of a hybrid approach that entails prior knowledge of the biological system and gene set enrichment analysis on cluster markers. An alternative, cluster-free approach to phenotypic identification of cellular states is trajectory analysis (Saelens et al., 2019), which aims to derive differentiation processes by using a pseudo-temporal ordering of single cells. However, in addition to identity- and differentiation-specific activities, transcriptional programs encompass a variety of cellular processes such as metabolism, growth, stress, and cell signaling, which are not necessarily captured by these approaches. Nevertheless, such expression programs are of great interest in a disease setting, where several communicating cell populations might act within the same dysregulated pathway. Thus, an in-depth characterization of such pathogenic signaling cascades at single-cell resolution is of great interest from a disease understanding perspective.

Latent factor models aim to decompose the global expression profile in its underlying transcriptional programs (Stein-O’Brien et al., 2018). These models project both genes and cells in a low-dimensional space, with latent dimensions approximating cells’ transcriptional programs and summarizing the contributions of several genes. Standard matrix factorization approaches, such as principal component analysis (PCA), non-negative matrix factorization (NMF), and independent component analysis (ICA), have been widely applied to scRNA-seq data (Kotliar et al., 2019). Nevertheless, novel methods have been developed that account

[1] Autoimmunity Transplantation and Inflammation Bioinformatics, Novartis Institutes for BioMedical Research, Novartis Campus, Basel 4056, Switzerland

[2] Present address: Helmholtz Zentrum München – German Research Center for Environmental Health, Institute of Computational Biology, 85764 Neuherberg, Germany

[3] Lead Contact

*Correspondence: enrico.ferrero@novartis.com

https://doi.org/10.1016/j.isci.2020.101451

for the specificities of single-cell data, using meaningful prior distributions and enforcing sparsity (Bielecki et al., 2018; Buettner et al., 2015; Levitin et al., 2019; Lopez et al., 2018; Pierson and Yau, 2015; Stein-O'Brien et al., 2019). A key parameter choice of these methods that is left to the analyst is the number of latent dimensions to use. Despite a few heuristics having been proposed based on stability analysis or model selection (Bielecki et al., 2018; Kotliar et al., 2019; Stein-O'Brien et al., 2019; Way et al., 2020), it is unclear whether these strategies could be applied effectively to datasets with different characteristics and whether such heuristics are appropriate for different downstream tasks. For example, it has been shown that different biological processes are captured at different dimensionalities of the latent space (Way et al., 2020), suggesting that approaches considering a varying number of latent dimensions could be more robust in fully recapitulating the underlying biology of the dataset under consideration.

To explore the potential of this family of methods to uncover previously unidentified pathway activities, we perform a systematic comparison of four recent latent factor models that specifically account for the sparsity of scRNA-seq data: scCoGAPS (Stein-O'Brien et al., 2019), LDA (Bielecki et al., 2018; Dey et al., 2017), scHPF (Levitin et al., 2019), and scVI (Lopez et al., 2018; Svensson et al., 2020). The first three methods are built on probabilistic approaches to matrix factorization and have been successfully used to extract gene signatures from scRNA-seq data (Clark et al., 2019; Svensson et al., 2020; Xu et al., 2019; Zhao et al., 2020), whereas the last one is based on a deep variational autoencoder with a linear decoder, making the inferred gene weights interpretable (Svensson et al., 2020). We test these on two scRNA-seq datasets from patients with autoimmune diseases. The first dataset consists of single cells isolated from synovial biopsies of patients with rheumatoid arthritis (RA) and sorted into four main cell subsets: monocytes, B cells, T cells, and fibroblasts (referred to as the RA dataset) (Zhang et al., 2019). The second dataset consists of single cells isolated from the kidney of patients with systemic lupus erythematosus (SLE) with lupus nephritis (LN) and enriched for the leukocyte component (referred to as the SLE dataset) (The Accelerating Medicines Partnership in SLE Network et al., 2019). We evaluate the stability over iterations of the four methods across the dimensionality of the latent space by using three different metrics and highlight the predictive power of these methods to discriminate cells isolated from patients or controls. Then, we assess the methods' ability to recover gene signatures by evaluating the coverage across 13 different gene set collections. Reasoning that latent factors can be used as surrogates of pathway activities, we devise a simple method to assign gene signatures to cell clusters, thus enabling the identification of cell subsets from a functional perspective. We then extend this analytical framework to integrate ligand – receptor interactions across cell subsets. Finally, we explore the reported gene signatures and discover two previously unidentified pathways in the RA dataset: the OSMR signaling pathway in a subpopulation of fibroblasts and the MERTK signaling pathway in a monocyte subset. We show that these signatures are potentially disease associated, thus highlighting the power of latent factor modeling to inform the discovery of novel pathogenic pathways.

RESULTS

Evaluation of Latent Factor Models Show Differences in Performance across Tasks and Latent Dimensions

It has been shown that factorization solutions are not strictly convex, thus resulting in different outputs for different iterations of the algorithms (Kotliar et al., 2019; Nordhausen, 2009). A common heuristics to select an appropriate latent dimension is to calculate the algorithm's stability across iterations and select the number of dimensions with results that are more consistent across iterations (Kotliar et al., 2019; Wu et al., 2016). For each method, we performed 10 iterations across 13 dimensionalities of the latent space (from k = 16 to k = 40, with step 2) and computed three stability metrics: Amari distance, silhouette score on the k-medoids-defined clusters, and the singular value canonical correlation analysis (SVCCA) score (Raghu et al., 2017) (see Methods for details). We performed this evaluation for both the RA and the SLE datasets (Figures 1A and 1B). scCoGAPS and LDA emerge as having better stability properties, across latent dimensions as well as across the three chosen metrics. In contrast, scVI shows poor stability, showing better performance than scHPF only for the SVCCA score. Overall, all methods report a lower performance as the number of latent dimension increases, consistent with the increased model complexity. End-of-training values for the objective functions of the four methods, at different values of k, were also investigated (Figures S1A and S1B).

To assess whether these latent factors retained information on the disease state of the samples, we used them as predictors of an elastic net regression model with the task of classifying disease and control cells

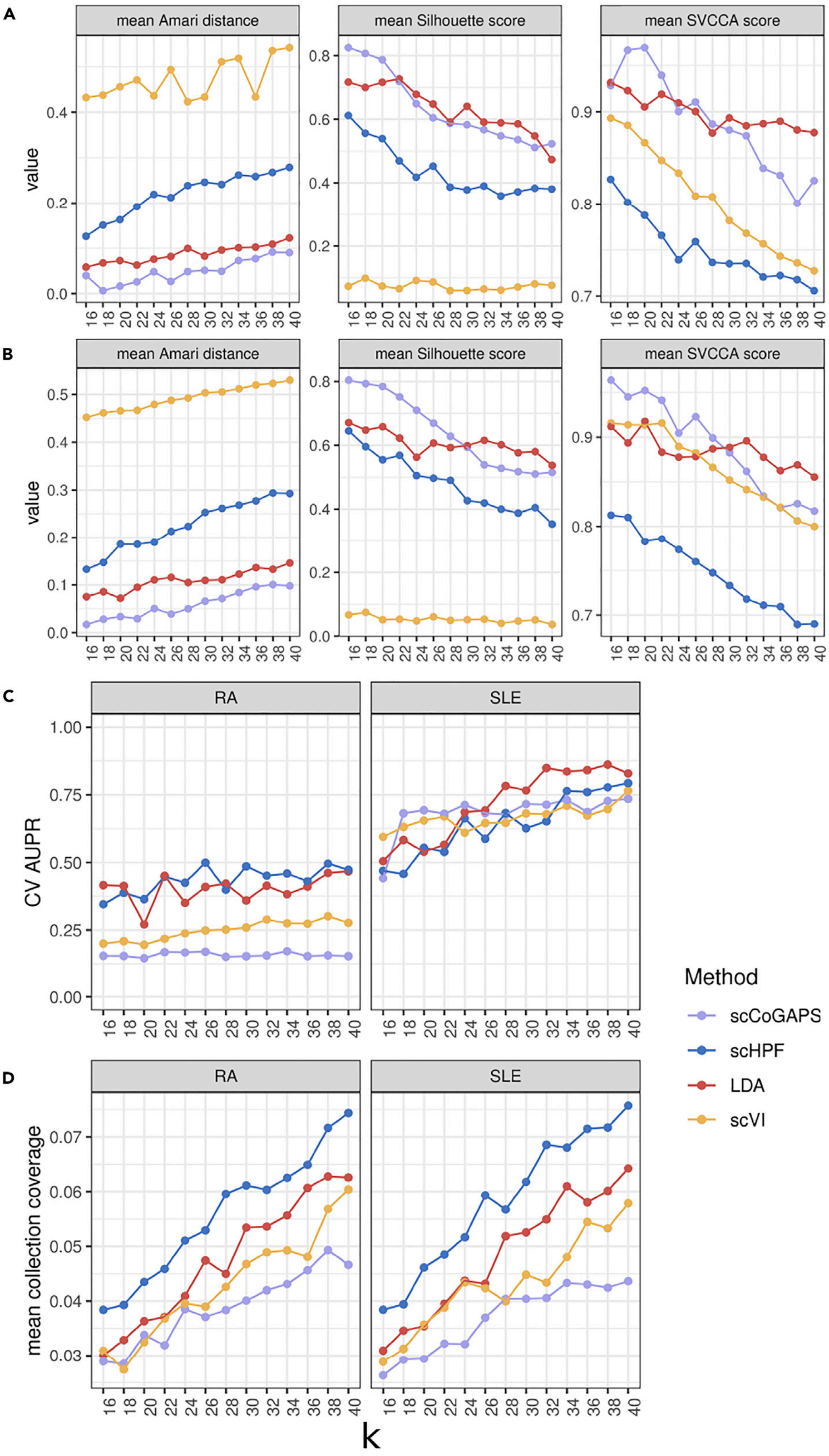

A
mean Amari distance
mean Silhouette score
mean SVCCA score
value
0.0
0.2
0.4
0.2
0.4
0.6
0.8
0.7
0.8
0.9
B
mean Amari distance
mean Silhouette score
mean SVCCA score
value
0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.2
0.4
0.6
0.8
0.7
0.8
0.9
C
RA
SLE
CV AUPR
0.00
0.25
0.50
0.75
1.00
D
RA
SLE
mean collection coverage
0.03
0.04
0.05
0.06
0.07
16 18 20 22 24 26 28 30 32 34 36 38 40
k
Method
scCoGAPS
scHPF
LDA
scVI

Figure 1. Evaluation of Latent Factor Models Show Differences in Performance across Tasks and Latent Dimensions

(A and B) Stability metrics in (A) the RA dataset and (B) the SLE dataset. Y axis reports the mean value of the metric across 10 iterations. X axis reports k, the number of latent dimensions.

(C) Cross-validation AUPR curve in a disease-control classification task using latent variables as predictors, in the RA dataset and the SLE dataset. Y axis reports cross-validation AUPR value across 10 iterations. X axis reports k, the number of latent dimensions.

(D) Mean gene set collection coverage across latent dimensions, in the RA dataset and the SLE dataset. Y axis reports mean collection coverage value, averaged across 13 gene set collections. X axis reports k, the number of latent dimensions.

(Figures 1C, S1C, and S1D). Interestingly, we found the performance to vary considerably between datasets, but not methods. In the RA dataset, almost all methods fail to reach an AUPR >0.5 regardless of the dimensionality of the latent space. In contrast, for the SLE dataset, all methods see a consistent increase in predictive power as the dimensionality increases, with scCoGAPS and LDA showing the best performance at low (16–24) and high (26–40) dimensionality of the latent space, respectively. Taken together, these results suggest that the dimensionality of the latent space is critical for extracting biological features related to the disease state of the cell.

The ability of latent factor models to recover biological signal is a key feature in their application to discover cellular phenotypes. Gene set enrichment analysis is a widely used approach for this task, as it allows mapping each latent variable to a specific pathway or biological process. To evaluate the methods' ability to recapitulate biologically meaningful gene signatures in a systematic manner, we used an enrichment approach based on heterogeneous network (Himmelstein et al., 2017; Way et al., 2020). Briefly, at each dimensionality of the latent space, we compute the gene set coverage score (number of unique gene sets significantly associated with each latent variable divided by the total number of gene set in the collection) for the gene set collection of interest (see Methods for details). We considered thirteen gene set collections, covering most of the known pathways and biological processes, as well as several other gene signatures (Figure 1D). As expected, for all methods we could observe an increase in the gene set coverage as the dimensionality of the latent space increases. By comparing the gene set coverage on the latent variables with the standard enrichment on clusters' marker genes (Figure S1E), we showed that the number of significant gene sets is an order of magnitude higher for the factorization methods, pointing to a higher sensitivity in the discovery of pathway activities. Interestingly, scHPF clearly outperformed the other methods in the majority of the gene set collections in both datasets (Figures S1F and S1G). This suggests that scHPF can decompose the expression matrix in a latent space that retains the highest degree of biological signal, which prompted us to use this method for all downstream analyses.

Systematic Assignment of Latent Variables to Cell Clusters Allows Identification of Cell Types Based on Their Phenotype or Function

An open challenge in single-cell transcriptomics is the phenotypic identification of cell populations after clustering. Usually, this is performed by means of a combination of prior knowledge of cell-specific markers and gene set enrichment analysis performed on the marker genes list for each cell subset (Luecken and Theis, 2019). However, as latent variables provide a surrogate of pathway activities across cells, we devised a simple framework to assign each pathway to cell clusters (Figure 2A). This approach directly allows the identification of cell subsets in a function- or phenotype-driven way. Briefly, we start by employing a standard clustering procedure to identify cell subsets (see Methods). For each gene set, we collapse redundant assignments to multiple latent variables in unique pathway activities, by means of an iterative clustering approach. Then, we regress pathway activity weights (i.e., the numerical results of the factorization) using the cell cluster labels as predictors. The coefficient of each cluster represents an indicator of how important that cell subset is to explain the pathway activity, therefore linking the activity of the pathway to the cell cluster, which can then be functionally interpreted. The heatmap in Figure 2B shows the coefficients for the most significant KEGG gene sets mapped to the latent variables obtained from the RA dataset, across the different cell clusters. Interestingly, broadly defined cell populations cluster together, showing that consistent activities across different biological processes recapitulate cell lineages. Importantly, this approach allows discovery of pathway activities that are unique to specific cell types or that are shared across different cell subsets in an unsupervised way. For instance, ''B cell signaling'' and ''NK cell cytotoxicity'' pathways (Figures 2C and 2D) show a distinct activity in the expected cell populations (see Figure S1H and S1I for an overview of the identified clusters). This strategy can also be used to annotate known, yet

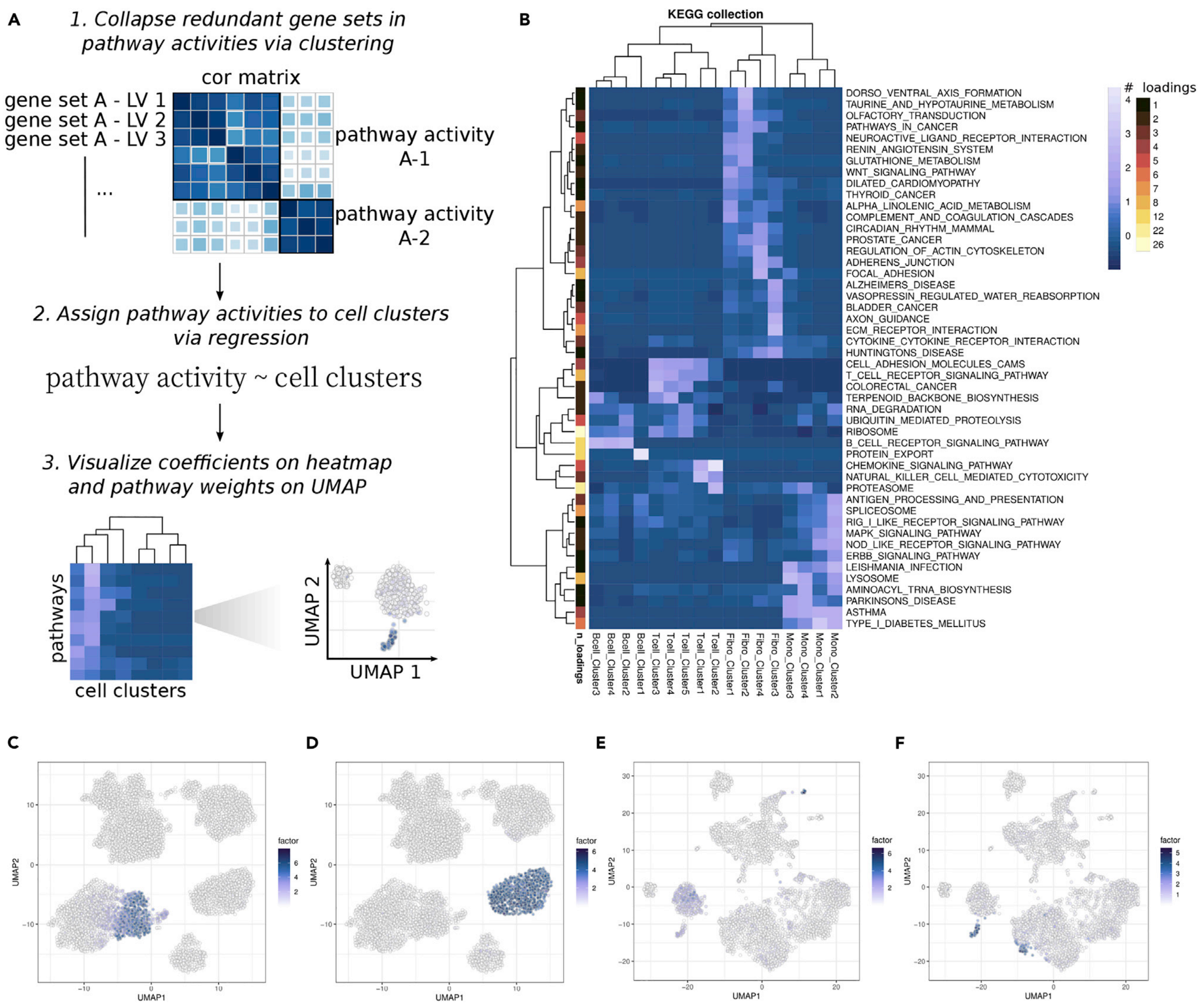


Figure 2. Systematic Assignment of Latent Variables to Cell Clusters Allows Identification of Cell Types Based on Their Phenotype or Function

(A) Schematic of the framework to derive pathway activities and assign them to cell subsets.

(B–F) (B) Heatmap of the KEGG collection's gene sets assigned to cell clusters in RA. The reported value represents the coefficient of the regression model and the "# loadings" color scale represents the number of latent variables that were found significant for that specific gene set. Factor weights for NK cell cytotoxicity gene set from the KEGG collection (C), B cell receptor signaling gene set from the KEGG collection (D), plasmacytoid dendritic cell signature from the C7 immunological signature collection (E), and type I interferon signaling gene set from the REACTOME collection (F).

unidentified, cell types, such as plasmacytoid dendritic cells (Figure 2E). Finally, in the SLE dataset, we could identify a type I interferon signature specifically active in a distinct subset of B cells and T cells, as previously reported (The Accelerating Medicines Partnership in SLE Network et al., 2019) (Figure 2F). Overall, this framework to define pathway activities is a powerful approach to assign cell identity and cell states to clusters based on their function or phenotype. Complete results for the RA and SLE datasets are reported in Figures S2 and S3 and Tables S1 and S2, respectively.

OSMR Signaling Is Active in Specific Subsets of Rheumatoid Arthritis Fibroblasts that Share a Similar Inflammatory Profile to Stromal Cells from Patients with Inflammatory Bowel Disease

By using latent variables as surrogates of pathway activities, we sought to discover novel pathways potentially involved in RA. We focused on Oncostatin M (OSM) receptor (OSMR) signaling, whose expression level was found to be low yet widespread across fibroblast subsets (Figure 3A). However, we found 17 latent factors enriched for OSMR signaling-related gene sets. We collapsed these redundant gene sets in four

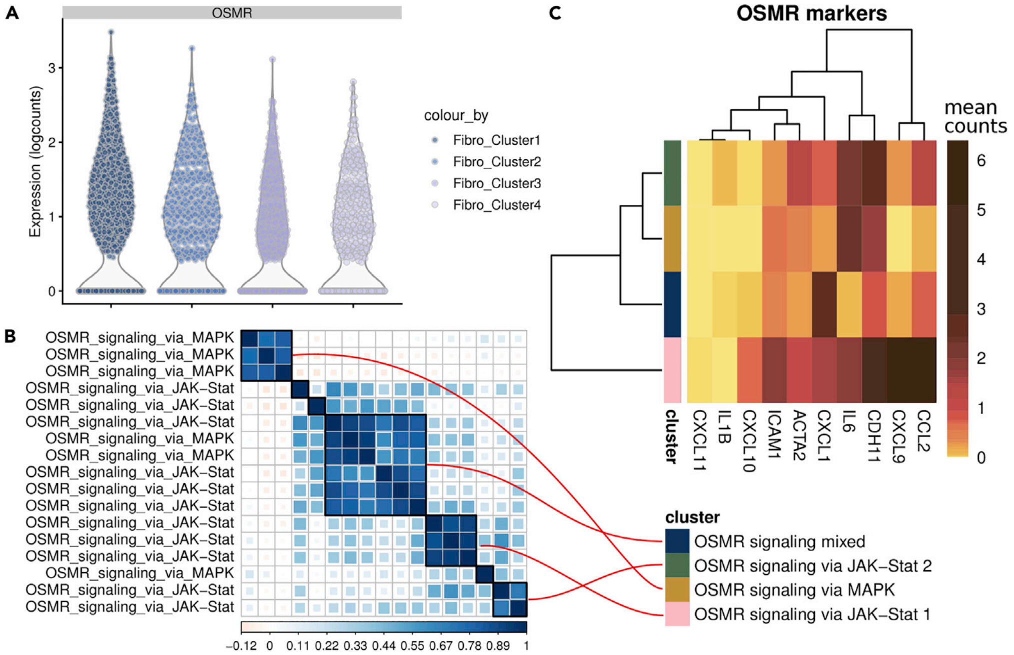


Figure 3. OSMR Signaling Is Active in Specific Subsets of Rheumatoid Arthritis Fibroblasts that Share a Similar Inflammatory Profile to Stromal Cells from Patients with Inflammatory Bowel Disease

(A) Expression levels of OSMR across fibroblast clusters.

(B) Correlation matrix of latent variables that maps to OSMR signaling pathways, as annotated by the METABASE gene set collection. Black frames enclose the correlation clusters that were selected to be representative of specific pathways activity.

(C) Mean expression level of genes found to be associated with OSMR-high-stromal cells in IBD.

pathway activities (Figure 3B), which showed a distinct distribution and composition of fibroblast subsets (Figures S4A and S4B). OSMR has been recently discovered to be a driver of increased inflammatory state of stromal cells in inflammatory bowel disease (IBD) (Oxford IBD Cohort Investigators et al., 2017). To investigate whether the RA fibroblast populations with high OSMR pathway activity also exhibited a similar inflammatory phenotype, we retrieved the gene signature associated with OSMR-high expression (Oxford IBD Cohort Investigators et al., 2017) and visualized the mean gene expression in the OSMR-signaling pathway activity (Figure 3C). Interestingly, two of the OSMR-related pathway activities showed a higher expression for the previously identified gene signature. Furthermore, these pathway activities seem to be mostly constituted by cells belonging to fibroblast clusters 1 and 2, which exhibit sublining markers (Figures S4C and S4D). These results indicate that OSMR signaling is active in synovial sublining fibroblasts in RA; as these cells share an inflammatory gene signature with OSMR-high stromal cells from patients with IBD, they suggest that OSMR could be a potential driver of inflammation also in RA.

Integration of Ligand-Receptor Interactions Reveals MERTK-Driven Apoptotic Cell Clearance by a Monocyte Subset in Rheumatoid Arthritis

To further explore the potential of pathway activities to uncover novel gene signatures, we sought to integrate this information with ligand-receptor interaction analysis (see Methods). In short, the expression level of interacting ligands and receptors was correlated with, and filtered for, latent variables with a significant enrichment for pathways where either protein was present (Figure 4A). Among the filtered cellular interactions, we found GAS6-MERTK. MERTK has a distinct expression across monocyte subsets (Figure 4B), and both monocyte clusters 1 and 3 showed interactions with GAS6 in fibroblasts and B cells subsets via MERTK (Figures S5A and S5B). We found MERTK-associated pathways to cluster in two main groups (Figure 4C): one with gene sets related to cell motility and cell signaling, the other one related to endocytosis and phagocytosis. As MERTK is a known marker for endocytic and phagocytic activity (particularly in the context of apoptotic cell clearance [Graham et al., 2014]), we set up to assess whether cells that were showing an endocytic-related activity showed an efferocytosis signature (Roberts et al., 2017; Waterborg et al., 2018). We observed that cells characterized by endocytic activity indeed recapitulated this gene signature to a

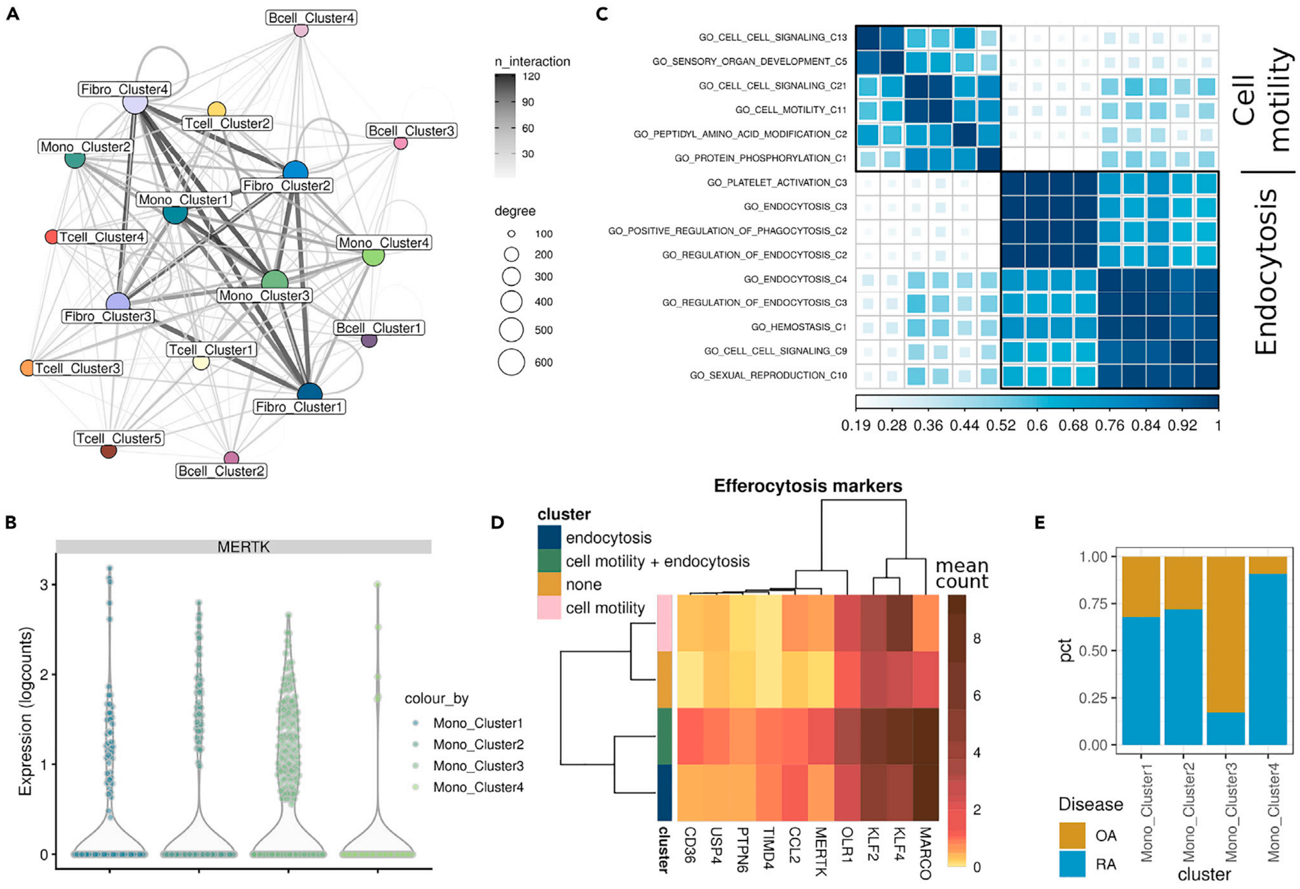


Figure 4. Integration of Ligand-Receptor Interactions Reveals MERTK-Driven Apoptotic Cell Clearance by a Monocyte Subset in Rheumatoid Arthritis

(A) Ligand-receptor interaction network as computed by CellPhoneDB and filtered as described in the main text.

(B) Expression levels of MERTK across monocyte clusters.

(C) Correlation matrix of latent factors that are associated with MERTK expression. Black frames enclose the correlation clusters that were selected to be representative of specific pathways activities.

(D) Mean expression levels of genes found to be associated with infiltrating macrophages showing an efferocytic activity.

(E) Proportion of cell types from disease (RA) or control (OA) across monocytes clusters.

higher degree as compared with the other activity clusters (Figure 4D). This cell subset, which is mostly constituted of monocytes from cluster 3 (Figure S5C, see Figure S5D for the signature enrichment across monocyte clusters), was found to be depleted in RA when compared with controls (Figure 4E), suggesting that reduced apoptotic cell clearance by MERTK-signaling monocytes could be a pathogenic mechanism in RA.

DISCUSSION

Latent factor models are a flexible approach to uncover transcriptional programs in an unsupervised fashion, thus allowing a functional annotation of cells in an unbiased way.

Here, we conducted an evaluation of four state-of-the-art latent factor models specifically developed for scRNA-seq data (scCoGAPS [Stein-O'Brien et al., 2019], LDA [Bielecki et al., 2018; Dey et al., 2017], scHPF [Levitin et al., 2019], and scVI [Lopez et al., 2018; Svensson et al., 2020]), assessing stability, predictive power, and gene set coverage of latent variables across the dimensionality of the latent space. Although the four methods represent considerably different approaches to the latent factor modeling paradigm, our evaluation highlighted some of the strengths and weaknesses of these methods across different tasks and can be used as a starting point for selecting the method of choice depending on the user needs.

We devised a novel framework to collapse redundant gene sets into pathway activities and assign these to cell clusters. We show that such an approach is able to retrieve known cellular phenotypes and that it can be used to identify cell subpopulations based on existing cell identity signatures, without having to rely on marker genes. Although we focused on two autoimmune disease cohorts, the described framework is generally applicable to any scRNA-seq datasets and provides an intuitive way to directly define cell subpopulations based on their function or phenotype.

Importantly, we also show that our framework can be used to discover previously unidentified pathways active in specific cell subsets. Among the pathways activities we retrieved that were not previously reported by the authors of the original publication (Zhang et al., 2019), we noticed the OSM signaling pathway. Interestingly, OSM has been recently reported to be a key driver of intestinal inflammation in IBD and to be associated with response to anti-TNF therapy (Oxford IBD Cohort Investigators et al., 2017). We showed that RA fibroblasts with an OSMR pathway activity also express higher levels of a gene signature associated with high expression of OSMR in IBD stromal cells. This points to the potential involvement of the OSM/OSMR axis in establishing the inflammatory microenvironment in patients with RA and suggests the pathway might be similarly dysregulated in RA and IBD. Of note, cells with increased OSMR signaling activity mainly belong to sublining fibroblast subsets. Since sublining fibroblasts have also been recently associated with a specific inflammatory phenotype (Croft et al., 2019) (as opposed to a more cartilage degradation phenotype of the lining fibroblasts), we can speculate that the OSM pathway is one of the drivers of inflammation not only in IBD but also in RA.

Through integration of ligand-receptor interaction analysis with our approach, we recovered a link between GAS6 on B cells and fibroblasts and MERTK in monocytes. As one of the pathway activities correlated to MERTK expression related to endocytosis, and because MERTK is known to be involved in efferocytosis, we hypothesized that MERTK-expressing monocytes could be involved in apoptotic cell clearance. To test this, we evaluated the expression of genes part of an efferocytosis signature (Roberts et al., 2017; Waterborg et al., 2018) and were able to show that this monocyte subset does indeed show expression of this signature and is depleted in disease. Interestingly, a recent report (Alivernini et al., 2020) confirmed the presence of MERTK$^+$ synovial macrophages driving remission in RA, thus substantiating our findings.

Limitations of the Study

We acknowledge an important limitation of our study is the relatively small selection of the latent factor models we pick for our initial evaluation. Our selection was guided by previous evidence that they could extract relevant transcriptional signatures from scRNA-seq data, as well as by the availability of implementations at the time of the analysis. Inclusion of other matrix factorization models such as f-scLVM (Buettner et al., 2017, 2015) and consensus NMF (Kotliar et al., 2019), or other deep learning approaches such as DCA (Eraslan et al., 2019), would certainly make any benchmarking efforts more comprehensive.

Another limitation is that both scRNA-seq datasets we analyzed were generated with the CEL-Seq2 technology and had a relatively low number of cells compared with more recently generated datasets (Martin et al., 2019; Schafflick et al., 2020; Smillie et al., 2019). The inclusion of larger datasets generated with different technologies would benefit a more comprehensive benchmarking effort and could potentially lead to different conclusions regarding the performance of the latent factor methods across different tasks.

Finally, as our approach to map latent factors to pathway activities and assign these to cell subsets relies heavily on gene sets annotation, the quality of the resulting pathway activities is inevitably tied to the quality of the original gene sets in the collection. As such, some of the identified pathway activities might be false positives. One example is the olfactory transduction gene set from the KEGG collection, which was found to be significantly associated with 64 latent factors, further collapsed in 19 pathway activities. Therefore, we suggest that hypothesis-driven explorations of the pathways activities assignments are needed to draw meaningful interpretations.

Resource Availability

Lead Contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Enrico Ferrero (enrico.ferrero@novartis.com).

Materials Availability

No materials were generated or used as part of this study.

Data and Code Availability

No new data was generated as part of this study. All code of the analysis is available at https://github.com/giovp/latent_factors_autoimmune.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.101451.

ACKNOWLEDGMENTS

We would like to thank Jonas Zierer, Christine Huppertz, Grigory Ryzhakov, Dominik Hartl, Stephan Spiegel, James Rush, and Richard Siegel for helpful discussions and comments on the manuscript and Patrick Dunn for helping with data access.

AUTHOR CONTRIBUTIONS

Conceptualization: E.F.; Methodology: E.F., G.P.; Formal Analysis: G.P.; Writing – Original Draft: G.P.; Writing – Review and Editing: E.F., G.P.; Visualization: G.P.; Supervision: E.F.

DECLARATION OF INTERESTS

E.F. is a Novartis employee and shareholder.

Received: November 27, 2019
Revised: May 28, 2020
Accepted: August 10, 2020
Published: September 25, 2020

REFERENCES

Alivernini, S., MacDonald, L., Elmesmari, A., Finlay, S., Tolusso, B., Gigante, M.R., Petricca, L., Di Mario, C., Bui, L., Perniola, S., et al. (2020). Distinct synovial tissue macrophage subsets regulate inflammation and remission in rheumatoid arthritis. Nat. Med. *26*, 1295–1306.

Bielecki, P., Riesenfeld, S.J., Kowalczyk, M.S., Vesely, M.C.A., Kroehling, L., Yaghoubi, P., Dionne, D., Jarret, A., Steach, H.R., McGee, H.M., et al. (2018). Skin inflammation driven by differentiation of quiescent tissue-resident ILCs into a spectrum of pathogenic effectors. bioRxiv.

Buettner, F., Natarajan, K.N., Casale, F.P., Proserpio, V., Scialdone, A., Theis, F.J., Teichmann, S.A., Marioni, J.C., and Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat. Biotechnol. *33*, 155–160.

Buettner, F., Pratanwanich, N., McCarthy, D.J., Marioni, J.C., and Stegle, O. (2017). f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. Genome Biol. *18*, 212.

Clark, B.S., Stein-O'Brien, G.L., Shiau, F., Cannon, G.H., Davis-Marcisak, E., Sherman, T., Santiago, C.P., Hoang, T.V., Rajaii, F., James-Esposito, R.E., et al. (2019). Single-cell RNA-seq analysis of retinal development identifies NFI factors as regulating mitotic exit and late-born cell specification. Neuron *102*, 1111–1126.e5.

Croft, A.P., Campos, J., Jansen, K., Turner, J.D., Marshall, J., Attar, M., Savary, L., Wehmeyer, C., Naylor, A.J., Kemble, S., et al. (2019). Distinct fibroblast subsets drive inflammation and damage in arthritis. Nature *570*, 246–251.

Crowell, H.L., Soneson, C., Germain, P.-L., Calini, D., Collin, L., Raposo, C., Malhotra, D., and Robinson, M.D. (2019). On the discovery of subpopulation-specific state transitions from multi-sample multi-condition single-cell RNA sequencing data. bioRxiv.

Der, E., Suryawanshi, H., Morozov, P., Kustagi, M., Goilav, B., Ranabothu, S., Izmirly, P., Clancy, R., Belmont, H.M., Koenigsberg, M., et al. (2019). Tubular cell and keratinocyte single-cell transcriptomics applied to lupus nephritis reveal type I IFN and fibrosis relevant pathways. Nat. Immunol. *20*, 915–927.

Dey, K.K., Hsiao, C.J., and Stephens, M. (2017). Visualizing the structure of RNA-seq expression data using grade of membership models. PLoS Genet. *13*, e1006599.

Eraslan, G., Simon, L.M., Mircea, M., Mueller, N.S., and Theis, F.J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. Nat. Commun. *10*, 1–14.

Fonseka, C.Y., Rao, D.A., Teslovich, N.C., Korsunsky, I., Hannes, S.K., Slowikowski, K., Gurish, M.F., Donlin, L.T., Lederer, J.A., Weinblatt, M.E., et al. (2018). Mixed-effects association of single cells identifies an expanded effector CD4$\mathplus$T cell subset in rheumatoid arthritis. Sci. Transl. Med. *10*, eaaq0305.

Graham, D.K., DeRyckere, D., Davies, K.D., and Earp, H.S. (2014). The TAM family: phosphatidylserine-sensing receptor tyrosine kinases gone awry in cancer. Nat. Rev. Cancer *14*, 769–785.

Himmelstein, D.S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S.L., Hadley, D., Green, A., Khankhanian, P., and Baranzini, S.E. (2017). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Elife *6*, e26726.

Keren-Shaul, H., Spinrad, A., Weiner, A., Matcovitch-Natan, O., Dvir-Szternfeld, R., Ulland, T.K., David, E., Baruch, K., Lara-Astaiso, D., Toth, B., et al. (2017). A unique microglia type Associated with restricting development of Alzheimer's disease. Cell *169*, 1276–1290.e17.

Kotliar, D., Veres, A., Nagy, M.A., Tabrizi, S., Hodis, E., Melton, D.A., and Sabeti, P.C. (2019). Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. Elife *8*, e43803.

Levitin, H.M., Yuan, J., Cheng, Y.L., Ruiz, F.J., Bush, E.C., Bruce, J.N., Canoll, P., Iavarone, A., Lasorella, A., Blei, D.M., and Sims, P.A. (2019). De novo gene signature identification from single-cell RNA -seq with hierarchical Poisson factorization. Mol. Syst. Biol. *15*, e8557.

Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat. Methods *15*, 1053–1058.

Luecken, M.D., and Theis, F.J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. Mol. Syst. Biol. *15*, e8746.

Ma, B.X., Korthauer, K., Kendziorski, C., and Newton, M.A. (2019). A compositional model to assess expression changes from single-cell Rna-seq data. bioRxiv.

Martin, J.C., Chang, C., Boschetti, G., Ungaro, R., Giri, M., Grout, J.A., Gettler, K., Chuang, L., Nayar, S., Greenstein, A.J., et al. (2019). Single-cell analysis of crohn's disease lesions identifies a pathogenic cellular module associated with resistance to anti-TNF therapy. Cell *178*, 1493–1508.e20.

Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J.Z., Menon, M., He, L., Abdurrob, F., Jiang, X., et al. (2019). Single-cell transcriptomic analysis of Alzheimer's disease. Nature *570*, 332–337.

Nordhausen, K. (2009). The elements of statistical learning: data mining, inference, and prediction, second edition by Trevor Hastie, robert Tibshirani, Jerome Friedman. Int. Stat. Rev. *77*, 482.

Oxford IBD Cohort Investigators, West, N.R., Hegazy, A.N., Owens, B.M.J., Bullers, S.J., Linggi, B., Buonocore, S., Coccia, M., Görtz, D., This, S., Stockenhuber, K., et al. (2017). Oncostatin M drives intestinal inflammation and predicts response to tumor necrosis factor–neutralizing therapy in patients with inflammatory bowel disease. Nat. Med. *23*, 579–589.

Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science *344*, 1396–1401.

Pierson, E., and Yau, C. (2015). ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. *16*, 241.

Puram, S.V., Tirosh, I., Parikh, A.S., Patel, A.P., Yizhak, K., Gillespie, S., Rodman, C., Luo, C.L., Mroz, E.A., Emerick, K.S., et al. (2017). Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. Cell *171*, 1611–1624.e24.

Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. (2017). SVCCA: singular vector canonical correlation analysis for deep learning dynamics and interpretability. arxiv.

Roberts, A.W., Lee, B.L., Deguine, J., John, S., Shlomchik, M.J., and Barton, G.M. (2017). Tissue-Resident macrophages are locally programmed for silent clearance of apoptotic cells. Immunity *47*, 913–927.e6.

Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. Nat. Biotechnol. *37*, 547–554.

Schafflick, D., Xu, C.A., Hartlehnert, M., Cole, M., Schulte-Mecklenbeck, A., Lautwein, T., Wolbert, J., Heming, M., Meuth, S.G., Kuhlmann, T., et al. (2020). Integrated single cell analysis of blood and cerebrospinal fluid leukocytes in multiple sclerosis. Nat. Commun. *11*, 247.

Smillie, C.S., Biton, M., Ordovas-Montanes, J., Sullivan, K.M., Burgin, G., Graham, D.B., Herbst, R.H., Rogel, N., Slyper, M., Waldman, J., et al. (2019). Intra- and inter-cellular rewiring of the human colon during ulcerative colitis. Cell *178*, 714–730.e22.

Soneson, C., and Robinson, M.D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. Nat. Methods *15*, 255–261.

Stein-O'Brien, G.L., Arora, R., Culhane, A.C., Favorov, A.V., Garmire, L.X., Greene, C.S., Goff, L.A., Li, Y., Ngom, A., Ochs, M.F., et al. (2018). Enter the matrix: factorization uncovers knowledge from omics. Trends Genet. *34*, 790–805.

Stein-O'Brien, G.L., Clark, B.S., Sherman, T., Zibetti, C., Hu, Q., Sealfon, R., Liu, S., Qian, J., Colantuoni, C., Blackshaw, S., et al. (2019). Decomposing cell identity for transfer learning across cellular measurements, platforms, tissues, and species. Cell Syst. *8*, 395–411.e8.

The Accelerating Medicines Partnership in SLE Network, Arazi, A., Rao, D.A., Berthier, C.C., Davidson, A., Liu, Y., Hoover, P.J., Chicoine, A., Eisenhaure, T.M., Jonsson, A.H., Li, S., et al. (2019). The immune cell landscape in kidneys of patients with lupus nephritis. Nat. Immunol. *20*, 902–914.

Svensson, V., Gayoso, A., Yosef, N., and Pachter, L. (2020). Interpretable factor models of single-cell RNA-seq via variational autoencoders. Bioinformatics *36*, 3418–3421.

Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science *352*, 189–196.

Vento-Tormo, R., Efremova, M., Botting, R.A., Turco, M.Y., Vento-Tormo, M., Meyer, K.B., Park, J.-E., Stephenson, E., Polański, K., Goncalves, A., et al. (2018). Single-cell reconstruction of the early maternal–fetal interface in humans. Nature *563*, 347–353.

Waterborg, C.E.J., Beermann, S., Broeren, M.G.A., Bennink, M.B., Koenders, M.I., van Lent, P.L.E.M., van den Berg, W.B., van der Kraan, P.M., and van de Loo, F.A.J. (2018). Protective role of the MER tyrosine kinase via efferocytosis in rheumatoid arthritis models. Front. Immunol. *9*, 742.

Way, G.P., Zietz, M., Rubinetti, V., Himmelstein, D.S., and Greene, C.S. (2020). Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations. Genome Biol. *21*, 109.

Wu, S., Joseph, A., Hammonds, A.S., Celniker, S.E., Yu, B., and Frise, E. (2016). Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks. Proc. Natl. Acad. Sci. U S A *113*, 4290–4295.

Xu, H., Ding, J., Porter, C.B.M., Wallrapp, A., Tabaka, M., Ma, S., Fu, S., Guo, X., Riesenfeld, S.J., Su, C., et al. (2019). Transcriptional Atlas of intestinal immune cells reveals that neuropeptide α-CGRP modulates group 2 innate lymphoid cell responses. Immunity *51*, 696–708.e9.

Yuan, D., Tao, Y., Chen, G., and Shi, T. (2019). Systematic expression analysis of ligand-receptor pairs reveals important cell-to-cell interactions inside glioma. Cell Commun. Signal. *17*, 48.

Zhang, F., Wei, K., Slowikowski, K., Fonseka, C.Y., Rao, D.A., Kelly, S., Goodman, S.M., Tabechian, D., Hughes, L.B., Salomon-Escoto, K., et al. (2019). Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry. Nat. Immunol. *20*, 928–942.

Zhao, W., Dovas, A., Spinazzi, E.F., Levitin, H.M., Upadhyayula, P., Sudhakar, T., Marie, T., Otten, M.L., Sisti, M., Bruce, J.N., et al. (2020). Deconvolution of cell type-specific drug responses in human tumor tissue with single-cell RNA-seq. bioRxiv.

Zhou, J.X., Taramelli, R., Pedrini, E., Knijnenburg, T., and Huang, S. (2017). Extracting intercellular signaling network of cancer tissues using ligand-receptor expression patterns from whole-tumor and single-cell transcriptomes. Sci. Rep. *7*, 8815.

iScience, Volume *23*

Supplemental Information

Latent Factor Modeling of scRNA-Seq Data Uncovers Dysregulated Pathways in Autoimmune Disease Patients

Giovanni Palla and Enrico Ferrero

Transparent Methods

Datasets

The two datasets we used are from the NIH Accelerating Medicine Partnership (AMP) consortium, and were accessed via Immport (https://www.immport.org/shared/home). The RA dataset (SDY998) (Zhang et al., 2019) consists of cells from 3 osteoarthritis (OA) patients and 22 rheumatoid arthritis (RA) patients, isolated from the synovium and further sorted via fluorescent activated cell sorting (FACS) into 4 subpopulations: fibroblasts, monocytes, B cells and T cells. The SLE dataset (SDY997) (the Accelerating Medicines Partnership in SLE network et al., 2019) consists of cells from 3 healthy donors and 15 LN patients, a complication of SLE which involves the kidney. Cells were isolated from the kidney and enriched for leukocytes by sorting for CD45+ cells.

Quality assessment, feature selection and clustering

scRNA-seq raw counts were downloaded from ImmPort. Quality assessment was performed with the *Scater* package (McCarthy et al., 2017) with the following thresholds. Cells had to have > 1000 and < 5000 UMI counts to filter out potential doublets as well as dead cells. The percentage of mitochondrial reads had to be below 0.25. Genes had to have at least 3 UMI in at least 3 cells. Further gene filtering was performed by means of the *deviance* (Townes et al., 2019) definition. In the end, 6000 and 3000 genes were selected for the RA and SLE datasets, respectively. Cell clustering was performed using the Louvain (Blondel et al., 2008) method applied to the shared nearest neighbour graph (obtained via the *scran* package (L. Lun et al., 2016), (Lun et al., 2016)), with the *clustree* package (Zappia and Oshlack, 2018) used to define clustering granularity. Clusters and labels from the original publications were not used in our analysis. UMAPs (Becht et al., 2018) of the two datasets and respective clusters are reported in Supplementary figure 1H and 1I.

Latent factor modelling algorithms

Four methods were selected for evaluation: scCoGAPS (Stein-O'Brien et al., 2019), LDA (Dey et al., 2017), scHPF (Levitin et al., 2019) and scVI (Lopez et al., 2018; Svensson and Pachter, 2019). The number of dimensions for the latent space (16 to 14 latent variables, step 2) has been selected for efficiency and for consistent evaluation across models. Here, the loading matrix refers to the matrix latent variables v. genes, whereas the factor matrix refers to the matrix cell v. latent variables. At each dimensionality of the latent space, we run the methods for 10 iterations. For scCoGAPS, we employed a parallelization approach as suggested by the authors, and set the maximum iteration parameter to 7000. For scVI we made use of the implementation of the linear decoder (Svensson and Pachter, 2019), so that the decoder weight matrix could be used as a surrogate of the loadings matrix and more easily interpreted. Model hyperparameters were set based on those selected for datasets of similar size and characteristics, as reported by the authors (Lopez et al., 2018). For LDA we employed hyperparameters as described in a similar use case (Bielecki et al., 2018). For scHPF, we employed hyperparameters as suggested by the authors (Levitin et al., 2019).

Stability evaluation

Stability evaluation was performed across iterations and latent dimensions. Three metrics were used:

- *Amari-like distance* (Hastie et al., 2017): a correlation-based metrics, that we computed for each pairs of iterations. The mean of all the comparisons is reported.

- *Silhouette score:* similar to Kotliar et al. (Kotliar et al., 2019), the silhouette score was calculated based on the clusters of the concatenated loading matrices. Briefly, ten loading matrices, computed for each iteration, were concatenated and further clustered with a k-medoids approach. The number of clusters was set equal to the number of latent dimensions. The mean of all the comparisons is reported.

- *SVCCA* (Raghu et al., 2017)*:* similar to Way et al. (Way et al., 2020) , SVCCA computes singular value decomposition on two loading matrices, and then perform canonical correlation analysis, to align matching components and derive correlation coefficients between them. The mean of all the comparisons is reported.

Classification

For each of the four methods, we used an elastic net logistic regression model with 10 fold cross-validation, using the latent factors as predictors and the disease state as the response variable. Given the unbalanced nature of the dataset we used AUPR as the evaluation metric.

Gene set coverage evaluation

Gene set coverage score was computed by means of heterogeneous networks (Himmelstein et al., n.d.), as described in Way et al (Way et al., 2020) . Briefly, we made use of heterogeneous network made available by Way et al. or generated as part of this study from several gene set collections (MSigDB (Liberzon et al., 2011), KEGG (Kanehisa and Goto, 2000), REACTOME (Jassal et al., 2020), Biocarta, MetaBase (Clarivate Analytics MetaBase® version 6.15.62452), WikiPathways (Slenter et al., 2018)). We also generated respective shuffled networks in order to calculate a z-score for each gene set – latent variable pair. We then converted the z-scores to p-values and applied a Bonferroni correction. Gene sets were considered significant if their p-value was lower than 0.01 divided by the number of latent variables for the specific latent space dimensionality. For each gene set collection and each latent variable, the top gene set was selected to be mapped to that latent variable. Ultimately, for each model, we would have at most an equal number of gene sets according to the dimensionality of the latent space. The number of unique gene sets was then used to calculate the coverage score (number of unique gene sets divided by the number of total gene sets in that collection). In Supplementary figure 1F and 1G, the mean collection coverage value was calculated across iterations for each dimensionality of the latent space, for the RA and SLE datasets, respectively. In Figure 1C, the mean collection coverage, across collections, was calculated for the best iteration at a fixed k, for each method. The selection of the best (i.e.: run with lowest reported loss) factorization result for a given k, was used for all downstream analyses. To perform the comparison with gene lists generated by means of differential expression across clusters, we assumed that each cluster-specific gene signature represent a latent feature. We used -log10 (p-value) as gene weights (analogous to the gene weights of the gene v. latent factors matrix that results from the matrix factorization step). The gene weights were then used as input for the gene set coverage evaluation. Since the number of clusters is 17, we showed a comparison of gene set coverage with the other methods at k=16 and k=18 (Supplementary figure 1E).

Gene set assignment to cell clusters

After matching each latent variable to its most significant gene set in each collection, we reasoned that we could use the latent variable as a surrogate of the respective pathway activity. Furthermore, we decided to use all the latent variables derived by the models, without selecting a single dimensionality of the latent space. For this and downstream analysis, we made use of the latent variables as computed by *scHPF,* selecting the best run (as assessed by the loss) for each dimensionality of the latent space. To account for the duplicated instances of the gene sets (same gene set mapped to multiple latent variables), we implemented a simple iterative algorithm. If a gene sets mapped to 3 or more latent variables, we would cluster the correlation matrix of the latent variables and compute the mean Silhouette width at different cut (H) of the hierarchical tree (where 1<H< # latent variables – 1). Then, for all the clusters that had a mean Silhouette width > 0.5 we collapsed them by computing the medians of the weights of the latent variables belonging to the respective clusters. If the Silhouette width <= 0.5 or the gene set mapped to only 2 latent variables, we just collapsed all the latent variables mapping to that gene set into the mean. This would provide us with collapsed latent variables for downstream analysis that we refer to as pathway activities. If the pathway activities were split during this clustering step, then it would be reported with a unique identifier (e.g. pathway_A_C2, representing the additional instance of pathway_A).

Furthermore, we devised an approach to assign these pathway activities to cell clusters: each pathway activity was set as the response variable in a regression setting where the cluster labels function as the predictors. Thus, a cell cluster that comprises several cells that have a high weight for that specific latent variable, would also be assigned a large coefficient. The pathway activities reported in Figure 2B represent the cell clusters' coefficients for all the pathway activities (see Supplementary figure 2 and 3 for the remaining collections). If the gene set was collapsed in several different pathway activities, those assignments are also reported.

Gene signature analysis

OSMR signature

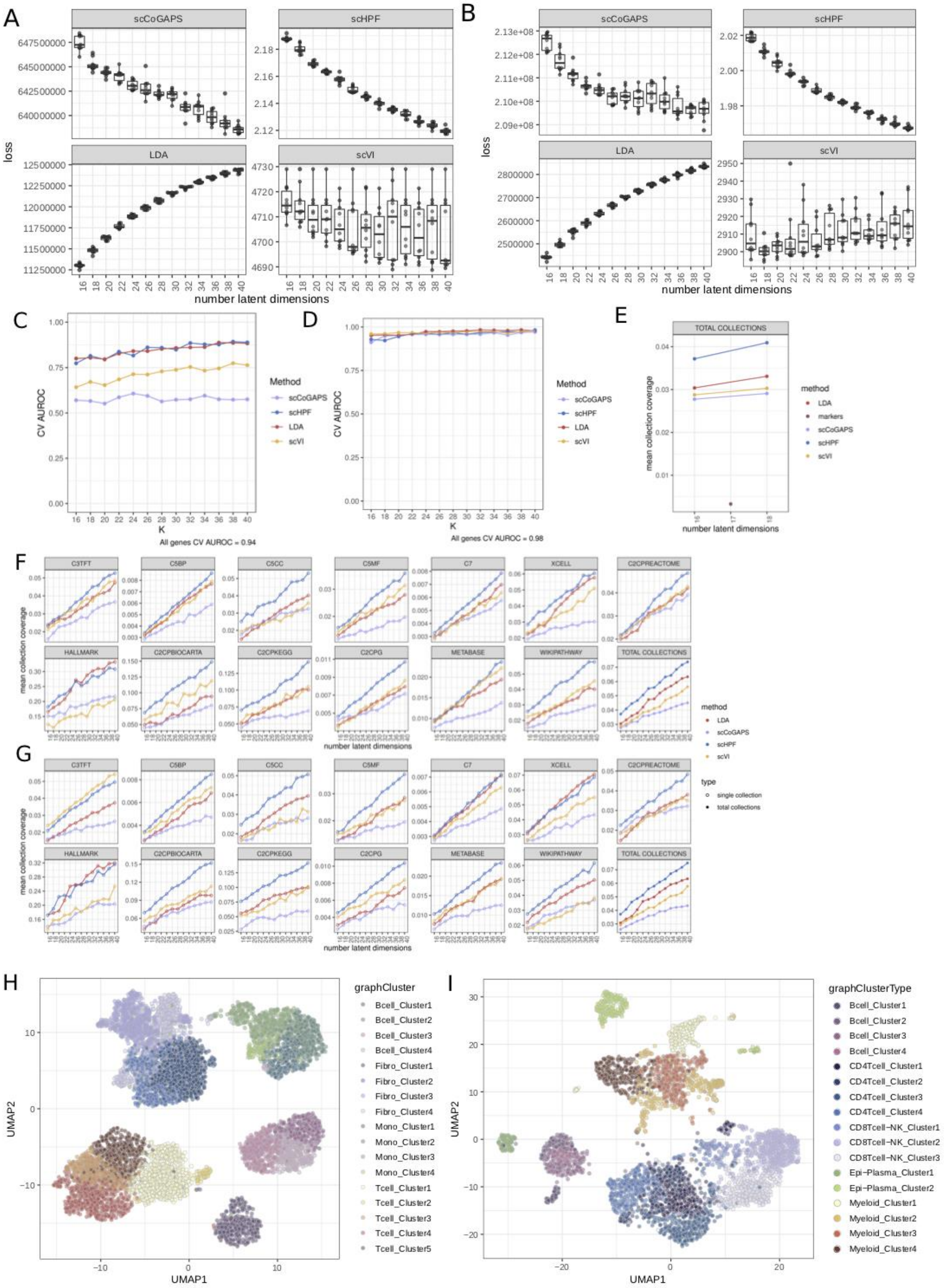
Pathways mapping to OSMR signalling were clustered and processed as described above. Only cells that showed a clear activity for the aforementioned pathway activities were retained (filtered based on matrix factorization weights above the median for all cells). Those cells consisted only of fibroblasts, and were used to visualize the gene expression signatures in Figure 3B. The mean gene expression signatures for the selected genes were calculated for cells that were specifically filtered in one of the four clusters but did not intersect with others. The expression of the same gene list was also reported for the fibroblast clusters (Supplementary figure 4C). The gene signature was retrieved from West et al (West et al., 2017).

MERTK signature

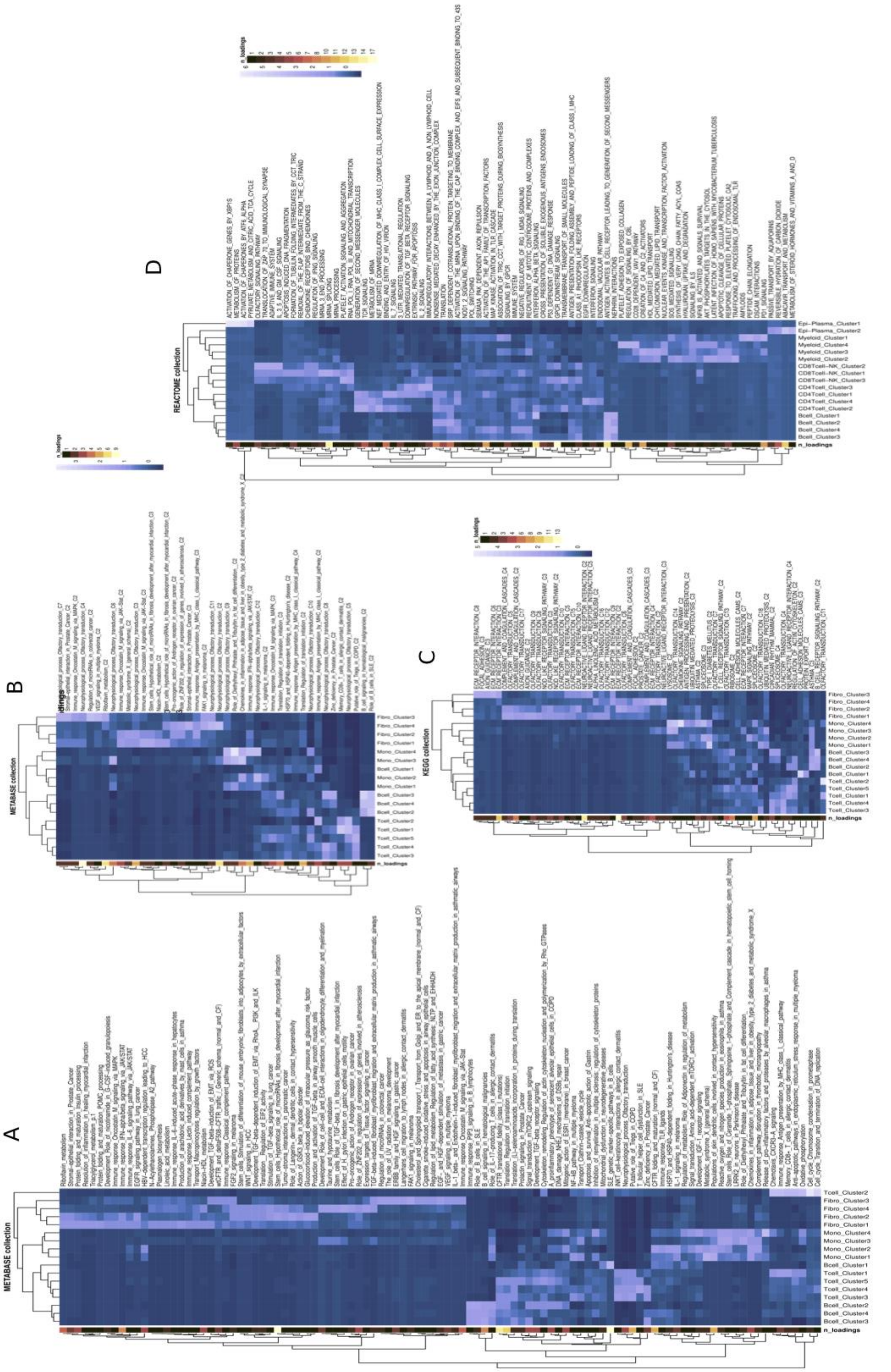
CellPhoneDB[17] was run with following parameters: *statistical_analysis, --iterations 5000, --threshold 0.2*. All gene sets activities (i.e.: weights of the latent variables) where MERTK is present were correlated with

MERTK expression levels, in monocyte cells. All gene sets that showed an R^2>0.3 were then filtered and the correlation matrix was clustered (Figure 4B). Medians of the latent factor weights were calculated for the two identified clusters, and cells were filtered for their respective activity (as described above). Cells that showed the activity for either, both or none of the two clusters were filtered and mean expression values were reported for genes retrieved from a gene from Waterborg et al. 2018 (Waterborg et al., 2018) and Robert et al. 2017 (Roberts et al., 2017). Supplementary figure 5D shows the mean expression of genes belonging to the efferocytosis signature based on clustering. Supplementary figure 5E shows the composition of the different pathway activities in terms of monocyte clusters.

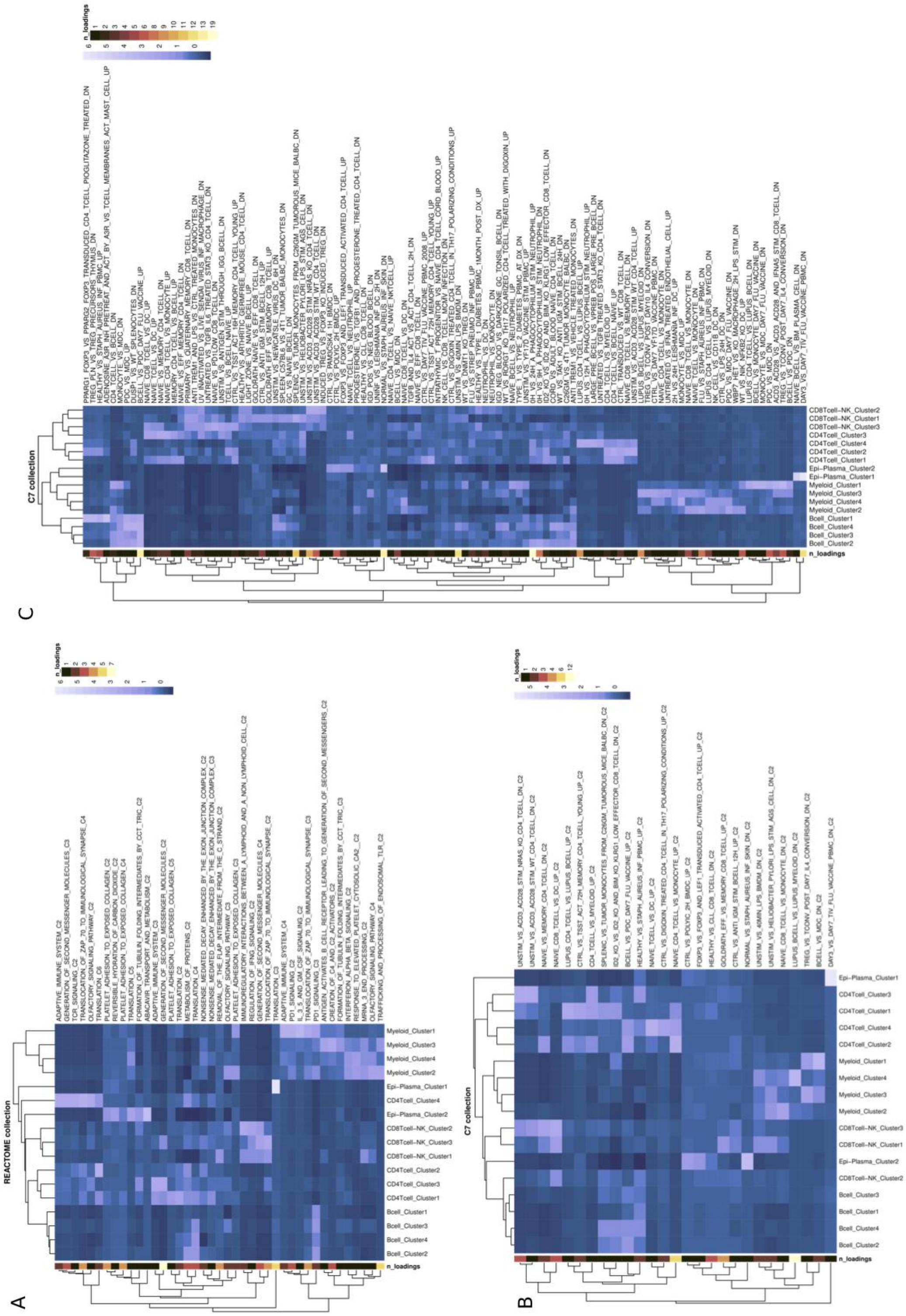
Supplementary Figures



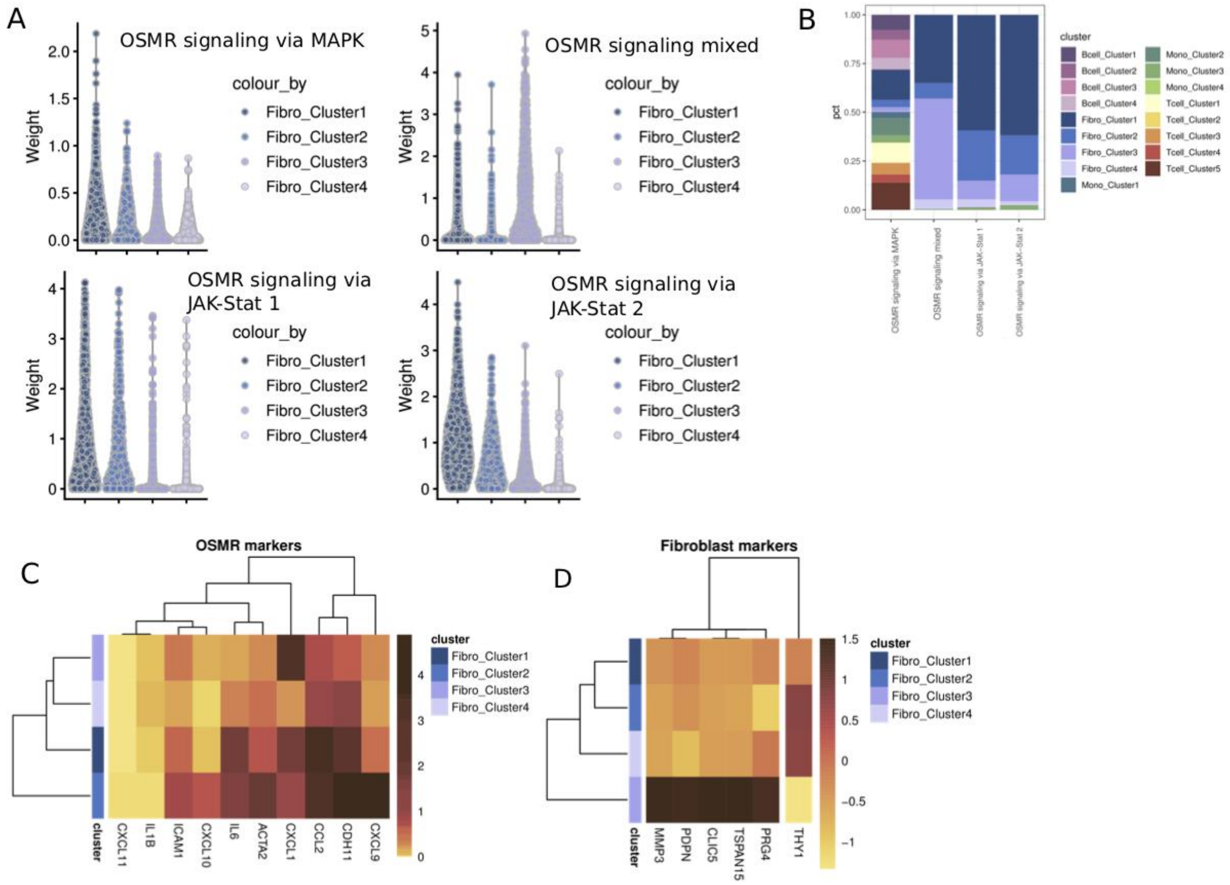
Supplementary figure 1: Value of the loss function used to monitor convergence after training across number of latent dimensions for the RA (A) and SLE (B) datasets; actual values are specific of the method (chi-square for scCoGAPS, Poisson likelihood for scHPF and scVI, log posterior probability for LDA). (C) AUROC across number of latent dimensions for the RA dataset and (D) for the SLE dataset. (E) Gene set coverage across gene set collections for the clusters markers as compared to the closest dimensionality of the latent space, for each method. Mean collection coverage value computed across iterations, at each dimensionality of the latent space and for each gene set collection, in the RA (F) and SLE (G) dataset. UMAP visualizing cell clusters in the RA (H) and SLE (I) dataset. Related to Figure 1.



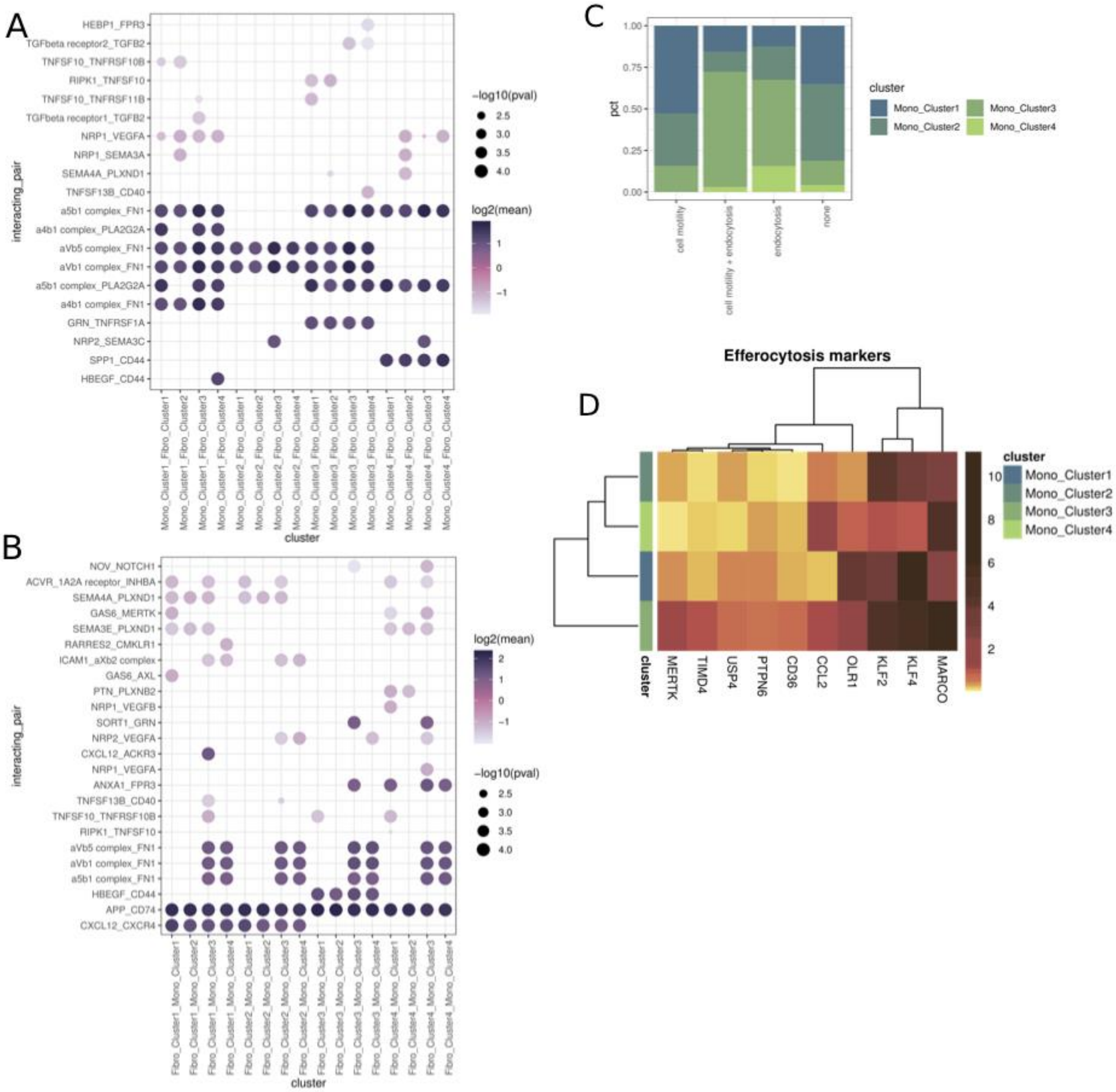
Supplementary figure 2: Heatmaps of gene sets assigned to cell clusters in the RA datasets for the METABASE (A, B), KEGG (C) and the REACTOME (D) collections. Related to Figure 2.



Supplementary figure 3: Heatmaps of gene sets assigned to cell clusters in the SLE dataset for the REACTOME (A) and the C7 (B, C) collections. Related to Figure 2.



Supplementary figure 4: (A) Weights of the four pathway activities clustered from the OSMR signalling-associated latent variables. (B) Composition of cell clusters in the respective pathway activities. (C) Expression of gene list described in Figure 3, across fibroblast clusters. (D) Marker genes for specific fibroblast subpopulations as described in Croft et al. 2019 **(Croft et al., 2019)**. Related to Figure 3.



Supplementary figure 5: CellPhoneDB protein-protein interaction analysis results, for (A) monocyte interacting with other cell subsets and (B) fibroblast interacting with other cell subsets. (C) Monocyte clusters composition of the identified pathway activities. (D) Efferocytosis gene signature across monocyte clusters. Related to Figure 4.

References (Transparent Methods)

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., Newell, E.W., 2018. Dimensionality reduction for visualizing single-cell data using UMAP. Nat. Biotechnol. 37, 38–44. https://doi.org/10.1038/nbt.4314

Bielecki, P., Riesenfeld, S.J., Kowalczyk, M.S., Vesely, M.C.A., Kroehling, L., Yaghoubi, P., Dionne, D., Jarret, A., Steach, H.R., McGee, H.M., Porter, C.B.M., Licona-Limon, P., Bailis, W., Jackson, R.P., Gagliani, N., Locksley, R.M., Regev, A., Flavell, R.A., 2018. Skin inflammation driven by differentiation of quiescent tissue-resident ILCs into a spectrum of pathogenic effectors. https://doi.org/10.1101/461228

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. 2008, P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Croft, A.P., Campos, J., Jansen, K., Turner, J.D., Marshall, J., Attar, M., Savary, L., Wehmeyer, C., Naylor, A.J., Kemble, S., Begum, J., Dürholz, K., Perlman, H., Barone, F., McGettrick, H.M., Fearon, D.T., Wei, K., Raychaudhuri, S., Korsunsky, I., Brenner, M.B., Coles, M., Sansom, S.N., Filer, A., Buckley, C.D., 2019. Distinct fibroblast subsets drive inflammation and damage in arthritis. Nature 570, 246–251. https://doi.org/10.1038/s41586-019-1263-7

Dey, K.K., Hsiao, C.J., Stephens, M., 2017. Visualizing the structure of RNA-seq expression data using grade of membership models. PLOS Genet. 13, e1006599. https://doi.org/10.1371/journal.pgen.1006599

Hastie, T., Tibshirani, R., Friedman, J.H., 2017. The elements of statistical learning: data mining, inference, and prediction, Second edition, corrected at 12th printing 2017. ed, Springer series in statistics. Springer, New York, NY.

Himmelstein, D.S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S.L., Hadley, D., Green, A., Khankhanian, P., Baranzini, S.E., n.d. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. eLife 6. https://doi.org/10.7554/eLife.26726

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., Loney, F., May, B., Milacic, M., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Weiser, J., Wu, G., Stein, L., Hermjakob, H., D'Eustachio, P., 2020. The reactome pathway knowledgebase. Nucleic Acids Res. 48, D498–D503. https://doi.org/10.1093/nar/gkz1031

Kanehisa, M., Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27–30. https://doi.org/10.1093/nar/28.1.27

Kotliar, D., Veres, A., Nagy, M.A., Tabrizi, S., Hodis, E., Melton, D.A., Sabeti, P.C., 2019. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. eLife 8, e43803. https://doi.org/10.7554/eLife.43803

L. Lun, A.T., Bach, K., Marioni, J.C., 2016. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. 17, 75. https://doi.org/10.1186/s13059-016-0947-7

Levitin, H.M., Yuan, J., Cheng, Y.L., Ruiz, F.J., Bush, E.C., Bruce, J.N., Canoll, P., Iavarone, A., Lasorella, A., Blei, D.M., Sims, P.A., 2019. *De novo* gene signature identification from single-cell RNA-seq with hierarchical Poisson factorization. Mol. Syst. Biol. 15, e8557. https://doi.org/10.15252/msb.20188557

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., Mesirov, J.P., 2011. Molecular signatures database (MSigDB) 3.0. Bioinformatics 27, 1739–1740. https://doi.org/10.1093/bioinformatics/btr260

Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., Yosef, N., 2018. Deep generative modeling for single-cell transcriptomics. Nat. Methods 15, 1053–1058. https://doi.org/10.1038/s41592-018-0229-2

Lun, A.T.L., McCarthy, D.J., Marioni, J.C., 2016. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Research 5, 2122. https://doi.org/10.12688/f1000research.9501.2

McCarthy, D.J., Campbell, K.R., Lun, A.T.L., Wills, Q.F., 2017. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics 33, 1179–1186. https://doi.org/10.1093/bioinformatics/btw777

Raghu, M., Gilmer, J., Yosinski, J., Sohl-Dickstein, J., 2017. SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability. ArXiv170605806 Cs Stat.

Roberts, A.W., Lee, B.L., Deguine, J., John, S., Shlomchik, M.J., Barton, G.M., 2017. Tissue-Resident Macrophages Are Locally Programmed for Silent Clearance of Apoptotic Cells. Immunity 47, 913-927.e6. https://doi.org/10.1016/j.immuni.2017.10.006

Slenter, D.N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., Mélius, J., Cirillo, E., Coort, S.L., Digles, D., Ehrhart, F., Giesbertz, P., Kalafati, M., Martens, M., Miller, R., Nishida, K., Rieswijk, L., Waagmeester, A., Eijssen, L.M.T., Evelo, C.T., Pico, A.R., Willighagen, E.L., 2018. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res. 46, D661–D667. https://doi.org/10.1093/nar/gkx1064

Stein-O'Brien, G.L., Clark, B.S., Sherman, T., Zibetti, C., Hu, Q., Sealfon, R., Liu, S., Qian, J., Colantuoni, C., Blackshaw, S., Goff, L.A., Fertig, E.J., 2019. Decomposing Cell Identity for Transfer Learning across Cellular Measurements, Platforms, Tissues, and Species. Cell Syst. 8, 395-411.e8. https://doi.org/10.1016/j.cels.2019.04.004

Svensson, V., Pachter, L., 2019. Interpretable factor models of single-cell RNA-seq via variational autoencoders. bioRxiv 737601. https://doi.org/10.1101/737601

the Accelerating Medicines Partnership in SLE network, Arazi, A., Rao, D.A., Berthier, C.C., Davidson, A., Liu, Y., Hoover, P.J., Chicoine, A., Eisenhaure, T.M., Jonsson, A.H., Li, S., Lieb, D.J., Zhang, F., Slowikowski, K., Browne, E.P., Noma, A., Sutherby, D., Steelman, S., Smilek, D.E., Tosta, P., Apruzzese, W., Massarotti, E., Dall'Era, M., Park, M., Kamen, D.L., Furie, R.A., Payan-Schober, F., Pendergraft, W.F., McInnis, E.A., Buyon, J.P., Petri, M.A., Putterman, C., Kalunian, K.C., Woodle, E.S., Lederer, J.A., Hildeman, D.A., Nusbaum, C., Raychaudhuri, S., Kretzler, M., Anolik, J.H., Brenner, M.B., Wofsy, D., Hacohen, N., Diamond, B., 2019. The immune cell landscape in kidneys of patients with lupus nephritis. Nat. Immunol. 20, 902–914. https://doi.org/10.1038/s41590-019-0398-x

Townes, F.W., Hicks, S.C., Aryee, M.J., Irizarry, R.A., 2019. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. Genome Biol. 20. https://doi.org/10.1186/s13059-019-1861-6

Waterborg, C.E.J., Beermann, S., Broeren, M.G.A., Bennink, M.B., Koenders, M.I., Lent, P.L.E.M. van, Berg, W.B. van den, Kraan, P.M. van der, Loo, F.A.J. van de, 2018. Protective Role of the MER Tyrosine Kinase via Efferocytosis in Rheumatoid Arthritis Models. Front. Immunol. 9. https://doi.org/10.3389/fimmu.2018.00742

Way, G.P., Zietz, M., Himmelstein, D.S., Greene, C.S., 2019. Sequential compression across latent space dimensions enhances gene expression signatures. bioRxiv 573782. https://doi.org/10.1101/573782

Way, G.P., Zietz, M., Rubinetti, V., Himmelstein, D.S., Greene, C.S., 2020. Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations. Genome Biol. 21, 109. https://doi.org/10.1186/s13059-020-02021-3

West, N.R., Hegazy, A.N., Owens, B.M.J., Bullers, S.J., Linggi, B., Buonocore, S., Coccia, M., Görtz, D., This, S., Stockenhuber, K., Pott, J., Friedrich, M., Ryzhakov, G., Baribaud, F., Brodmerkel, C., Cieluch, C., Rahman, N., Müller-Newen, G., Owens, R.J., Kühl, A.A., Maloy, K.J., Plevy, S.E., Oxford IBD Cohort Investigators, Arancibia, C., Bailey, A., Barnes, E., Bird-Lieberman, B., Brain, O., Braden, B., Collier, J., East, J., Howarth, L., Keshav, S., Klenerman, P., Leedham, S., Palmer, R., Powrie, F., Rodrigues, A., Simmons, A., Sullivan, P., Travis, S.P.L., Uhlig, H., Keshav, S., Travis, S.P.L., Powrie, F., 2017. Oncostatin M drives intestinal inflammation and predicts response to tumor necrosis factor–neutralizing therapy in patients with inflammatory bowel disease. Nat. Med. 23, 579–589. https://doi.org/10.1038/nm.4307

Zappia, L., Oshlack, A., 2018. Clustering trees: a visualization for evaluating clusterings at multiple resolutions. GigaScience 7. https://doi.org/10.1093/gigascience/giy083

Zhang, F., Wei, and K., Slowikowski, K., Fonseka, C.Y., Rao, D.A., Kelly, S., Goodman, S.M., Tabechian, D., Hughes, L.B., Salomon-Escoto, K., Watts, G.F.M., Jonsson, A.H., Rangel-Moreno, J., Meednu, N., Rozo, C., Apruzzese, W., Eisenhaure, T.M., Lieb, D.J., Boyle, D.L., Mandelin, A.M., Boyce, B.F., DiCarlo, E., Gravallese, E.M., Gregersen, P.K., Moreland, L., Firestein, G.S., Hacohen, N., Nusbaum, C., Lederer, J.A., Perlman, H., Pitzalis, C., Filer, A., Holers, V.M., Bykerk, V.P., Donlin, L.T., Anolik, J.H., Brenner, M.B., Raychaudhuri, S., 2019. Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry. Nat. Immunol. 20, 928–942. https://doi.org/10.1038/s41590-019-0378-1