Tensor decomposition reveals coordinated multicellular patterns of transcriptional variation that distinguish and stratify disease individuals.

Jonathan Mitchel[1,2], M. Grace Gordon[3,5], Richard K. Perez[4], Evan Biederstedt[1], Raymund Bueno[3,5], Chun Jimmie Ye[3§], Peter V. Kharchenko[1,6§]

[1] Department of Biomedical Informatics, Harvard Medical School, Boston, MA
[2] Program in Health Sciences & Technology, Harvard Medical School & Massachusetts Institute of Technology, Boston, MA
[3] Institute for Human Genetics, University of California, San Francisco, San Francisco, CA
[4] School of Medicine, University of California, San Francisco, CA
[5] UCSF Division of Rheumatology, Department of Medicine, University of California, San Francisco, CA
[6] Broad Institute, Cambridge, MA

[§] Correspondence should be addressed to peter.kharchenko@post.harvard.edu and jimmie.ye@ucsf.edu

Summary

Tissue- and organism-level biological processes often involve coordinated action of multiple distinct cell types. Current computational methods for the analysis of single-cell RNA-sequencing (scRNA-seq) data, however, are not designed to capture co-variation of cell states across samples, in part due to the low number of biological samples in most scRNA-seq datasets. Recent advances in sample multiplexing have enabled population-scale scRNA-seq measurements of tens to hundreds of samples. To take advantage of such datasets, here we introduce a computational approach called single-cell Interpretable Tensor Decomposition (scITD). This method extracts "multicellular gene expression patterns" that vary across different biological samples. These patterns capture how changes in one cell type are connected to changes in other cell types. The multicellular patterns can be further associated with known covariates (e.g., disease, treatment, or technical batch effects) and used to stratify heterogeneous samples. We first validated the performance of scITD using *in vitro* experimental data and simulations. We then applied scITD to scRNA-seq data on peripheral blood mononuclear cells (PBMCs) from 115 patients with systemic lupus erythematosus and 56 healthy controls. We recapitulated a well-established pan-cell-type signature of interferon-signaling that was associated with the presence of anti-dsDNA autoantibodies and a disease activity index. We further identified a novel multicellular pattern that appears to potentiate renal involvement for patients with anti-dsDNA autoantibodies. This pattern was characterized by an expansion of activated memory B cells along with helper T cell activation and is predicted to be mediated by an increase in ICOSLG-ICOS interaction between monocytes and helper T cells. Finally, we applied scITD to two PBMC datasets from patients with COVID-19 and identified reproducible multicellular patterns that stratify patients by disease severity. Overall, scITD is a flexible method for exploring co-variation of cell states in multi-sample single-cell datasets, which can yield new insights into complex non-cell-autonomous dependencies that define and stratify disease.

Introduction

Gene expression is a defining feature that distinguishes different cell types. However, gene expression profiles derived from the same cell type can also vary across individuals, driven by a combination of genetics and environment. Most often, gene expression is compared between individuals in case-control studies or is used to infer sample subgroupings. Alternatively, studies such as the Genotype-Tissue Expression (GTEx) project have revealed the genetic basis of tissue-specific gene expression by mapping natural genetic variation associated with expression differences in human populations (Consortium, 2015; Melé et al., 2015). Applications of single-cell RNA-sequencing (scRNA-seq) have so far focused on the characterization of transcriptional differences among different cell types and cell states, and analysis of inter-individual variation has been hampered by small sample sizes and the presence of technical batch effects that are often difficult to separate from biological variation. Experimentally, this stimulated the development of multiplexed designs where samples from multiple individuals could be profiled in one run, thereby reducing confounding by technical batches (Kang et al., 2018; Stoeckius et al., 2018). Analytically, a variety of approaches have been developed to perform dataset alignment (Butler et al., 2018; Haghverdi et al., 2018; Barkas et al., 2019; Stuart et al., 2019). These tools establish correspondence between samples, effectively treating the difference between individuals as a problem to be overcome. These differences (e.g., between cases and controls or genetically different individuals) are often central for the downstream biological interpretation, but relatively few approaches exist to explore them systematically.

The existing methods to compare single-cell transcriptomes between individuals generally require pre-defined groups of samples such as in case/control studies. Comparisons between sample groups can be made using cell-level data or aggregated counts for all cells within a given type or cluster (often referred to as a "pseudobulk" operation) (Chen et al., 2020; Crowell et al., 2020).

After computing the pseudobulk profiles, differential expression (DE) tools designed for bulk RNA-seq analysis can be used to compare sample groups one cell type/cluster at a time. Pseudobulk analysis has been demonstrated to provide a superior balance of robustness, performance, and runtime (Crowell et al., 2020; Squair et al., 2021) and has been used in several single-cell case-control studies to date (Kang et al., 2018; Mathys et al., 2019; Corridoni et al., 2020; Liu et al., 2021; Ren et al., 2021; van der Wijst et al., 2021).

However, case-control DE approaches are unable to stratify patients into subgroups and do not account for expression dysregulation that may occur jointly in multiple cell types. Patient subgrouping is of interest when the covariate defining a group has not been well-captured (e.g., disease status has been only partially recorded or recorded with errors) or when additional heterogeneity across samples exists. There is a lack of unsupervised single-cell analysis methods designed for capturing inter-individual variation. Standard matrix decomposition methods such as principal components analysis (PCA) and non-negative matrix factorization can, in principle, be used to describe gene expression variation across pseudobulk samples one cell type/cluster at a time. However, as we will illustrate, it is more informative to consider inter-individual variation within multiple cell populations jointly. Such a joint decomposition would more naturally describe scenarios where different cell types respond specifically to the same external signals. It would also improve our ability to infer dependencies between transcriptional programs across cell types (e.g., due to cell-cell communication or interaction) (Browaeys et al., 2020; Cabello-Aguilar et al., 2020; Efremova et al., 2020; Jin et al., 2021). For example, a sample undergoing an innate immune response may display increased chemokine expression in myeloid cells and increased expression of chemotaxis genes in neutrophils, as these processes tend to occur together. Overall, by extracting such patterns of gene expression variation, we hypothesize that we can better characterize the molecular bases of complex phenotypes.

Here, we developed an unsupervised computational method, called single-cell Interpretable Tensor Decomposition (scITD), that can infer multicellular patterns of gene expression (Figure 1A). We define a “multicellular pattern” to be a collection of genes in various cell types that co-vary together across samples. The multicellular patterns inferred by scITD can be linked with various clinical annotations, genetics, technical batch effects, and other sample metadata, leading to a richer understanding of the system under study.

We first assessed the performance of scITD using simulated and real data from an *in vitro* experiment. Then, we applied scITD to investigate inter-individual heterogeneity in peripheral blood mononuclear cell (PBMC) expression using a dataset with 115 systemic lupus erythematosus (SLE) patients and 56 healthy controls. SLE is a heterogeneous autoimmune disease that can manifest with a wide array of symptoms and has few available targeted therapies (Fava and Petri, 2019; Allen et al., 2021). We identified six multicellular patterns of gene expression that stratify SLE patients, and we show that these are associated with clinical variables including disease activity and nephritis, one of the most severe complications of SLE. These patterns were examined in depth to identify channels of intercellular communication and changes in cell-type composition. We also applied scITD to a PBMC dataset consisting of 83 patients with COVID-19 and 20 healthy controls, revealing multicellular patterns associated with disease severity. These were validated in an independent study of 49 COVID-19 patients and 11 controls, pointing to conserved mechanisms in IL-16 signaling that could lead to new therapeutic opportunities. Finally, we compared multicellular patterns from the COVID-19 dataset to those from the SLE dataset, revealing similarities and differences in type-1 interferon-stimulated gene response that predispose individuals to autoimmunity but may be protective in acute viral infection.

Results

Approach and evaluation of performance

To extract multicellular patterns that vary across individuals, we first generate normalized pseudobulk expression profiles per sample per cell type (Methods). When C cell populations from N samples are collapsed into pseudobulk profiles, the dataset can be represented as a 3-dimensional matrix – a tensor T with dimensions $N \times G \times C$, where G is the number of genes (Figure 1B left). Key to scITD, to capture recurrent patterns of transcriptional variation across individuals, we applied Tucker tensor decomposition (Tucker, 1966) to extract the K most informative factors. In this context, each factor consists of two elements (Figure 1A middle). The first element is a gene-by-cell type matrix of loadings values, representing a multicellular pattern (Figure 1B right and 1E right), and the second element is a vector of sample scores indicating the relative amount of the multicellular pattern present in each sample (Figure 1B middle and 1E left). Therefore, the output structure for a decomposition to K factors consists of a factor-by-gene-by-cell type ($K \times G \times S$) tensor (Figure 1B right) and a sample-by-factor ($N \times K$) matrix (Figure 1B middle). We will refer to the former as the "loadings tensor" and the latter as the "sample scores matrix" or "donor scores matrix" (if there is only one sample per donor). These two data objects can be multiplied together to reconstruct an approximation of the original gene expression tensor, T. This representation has a notable advantage, in that each horizontal slice of the loadings tensor represents a multicellular pattern of gene expression that varies across samples according to the corresponding sample scores.

We first demonstrate scITD by applying it to a dataset with one primary driving source of variation across samples. This dataset consists of 16 samples of PBMCs from SLE patients, half of which were stimulated *in vitro* with interferon-beta (IFN-beta) (Kang et al., 2018) (Figure 1C). For simplicity of the demonstration, we limited the analysis to just classical monocytes (cMonocytes) and CD4+ T cells (Th cells). Differential expression (DE) between stimulated and control samples revealed both shared and cell-type-specific gene expression changes (Figure 1D). For example, the gene *MX1* becomes upregulated in both cell types after IFN-beta stimulation, whereas *ANXA5* becomes upregulated specifically in the monocytes. After applying scITD to this dataset, we examined the sample scores and loadings matrix for the first factor (Figure 1E). The scores for this factor perfectly separate the IFN-beta stimulated samples from the control samples (Figure 1E left). The corresponding factor slice of the loadings tensor reveals a similar pattern with both shared and cell-type-specific perturbed genes (Figure 1E right). Next, we computed associations between sample scores of factor 1 and the expression of each gene in each cell type to determine the genes with statistically significant contributions in the loadings matrix. The Benjamini-Hochberg (BH) procedure was used for multiple hypothesis test correction here as well as for other p-value adjustments throughout the study (Benjamini and Hochberg, 1995). The gene expression-factor associations showed a high concordance with the p-values obtained from the regular DE analysis (Figure 1F). Similarly, we observed a high correlation between the factor loadings for all genes and the log fold-change values from the DE analysis (Figure 1G). This basic example demonstrates that scITD can accurately extract multicellular expression patterns that involve both shared and cell-type-specific genes.

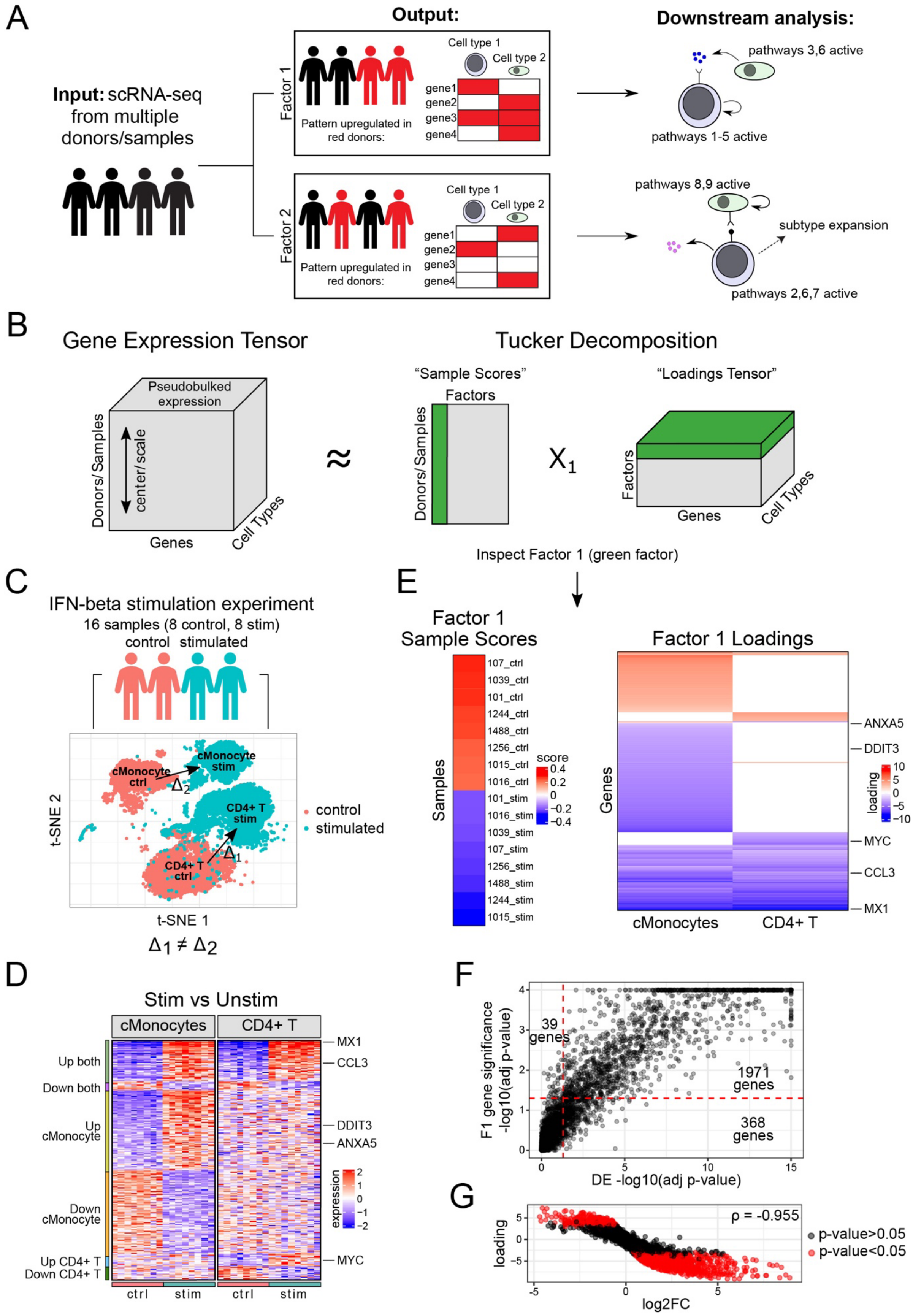

A
Input: scRNA-seq from multiple donors/samples
Output:
Factor 1
Cell type 1
Cell type 2
gene1
gene2
gene3
gene4
Pattern upregulated in red donors:
Factor 2
Cell type 1
Cell type 2
gene1
gene2
gene3
gene4
Pattern upregulated in red donors:
Downstream analysis:
pathways 3,6 active
pathways 1-5 active
pathways 8,9 active
subtype expansion
pathways 2,6,7 active
B
Gene Expression Tensor
Pseudobulked expression
Donors/Samples
center/scale
Genes
Cell Types
≈
Tucker Decomposition
"Sample Scores"
Factors
Donors/Samples
X1
"Loadings Tensor"
Factors
Genes
Cell Types
Inspect Factor 1 (green factor)
C
IFN-beta stimulation experiment
16 samples (8 control, 8 stim)
control stimulated
cMonocyte ctrl
cMonocyte stim
Δ2
CD4+ T stim
CD4+ T ctrl
Δ1
t-SNE 2
t-SNE 1
control
stimulated
Δ1 ≠ Δ2
D
Stim vs Unstim
cMonocytes
CD4+ T
Up both
Down both
Up cMonocyte
Down cMonocyte
Up CD4+ T
Down CD4+ T
MX1
CCL3
DDIT3
ANXA5
MYC
expression
2
1
0
−1
−2
ctrl
stim
ctrl
stim
E
Factor 1 Sample Scores
Samples
107_ctrl
1039_ctrl
101_ctrl
1244_ctrl
1488_ctrl
1256_ctrl
1015_ctrl
1016_ctrl
101_stim
1016_stim
1039_stim
107_stim
1256_stim
1488_stim
1244_stim
1015_stim
score
0.4
0.2
0
−0.2
−0.4
Factor 1 Loadings
Genes
ANXA5
DDIT3
MYC
CCL3
MX1
loading
10
5
0
−5
−10
cMonocytes
CD4+ T
F
F1 gene significance -log10(adj p-value)
DE -log10(adj p-value)
39 genes
1971 genes
368 genes
0
1
2
3
4
0
5
10
15
G
ρ = -0.955
loading
log2FC
5
0
-5
-5
0
5
p-value>0.05
p-value<0.05

Figure 1. General overview of scITD and demonstration of functionality.

(A) The overall goal of scITD. The tool takes clustered and annotated scRNA-seq data from multiple samples/donors as input (left). scITD then identifies multicellular patterns of gene expression that vary across the samples (middle). These patterns can be further analyzed to reveal biological processes that are jointly active in multiple cell types (right).

(B) Structure of the output from scITD (middle and right) applied to a single-cell pseudobulk expression tensor (left). An approximation of the expression tensor is reconstructed when the sample scores matrix (middle) and loadings tensor (right) are multiplied together. Sample scores and loadings for one factor are highlighted in green.

(C) scRNA-seq data from an IFN-beta stimulation experiment and t-SNE plot with cells colored by their corresponding sample stimulation condition.

(D) Sample-level pseudobulk gene expression of DE genes between control and IFN-beta stimulated samples. Rows are genes and columns are pseudobulked samples. Genes that are significant in at least one of the two cell types below an adjusted p-value of 0.01 were included. Genes are grouped (left annotation) by the cell types where they were DE across conditions. A few DE genes are shown labeled on the right.

(E) The sample scores (left) and loadings (right) for factor 1 after applying scITD to the IFN-beta stimulation data. Samples in the sample scores vector are labeled by their condition. Only loadings for significant genes in each of the two cell types are shown in the loadings heatmap (Methods). Rows of the loadings heatmap are hierarchically clustered. The same DE gene callouts from (D) are shown as labels on the right.

(F) Comparison of DE adjusted p-values (from D) to gene expression-factor 1 sample score association p-values (Methods). Each point is a gene in one of the two cell types. The dashed red lines are located at an adjusted p-value of 0.05.

(G) Comparison of DE log2 fold-change values (from D) to loadings values from factor 1. Each point is a gene in one of the two cell types. The Spearman correlation is shown in the upper-right corner. Red dots represent genes that are significantly associated with factor 1 at an adjusted p-value < 0.05.

To further evaluate the performance of scITD, we simulated a scRNA-seq dataset with 40 donors (1 sample per donor) and two cell types (Figure S1A left). The dataset was designed to include two multicellular patterns that varied across donors (Figure S1A right), and each pattern involved mostly different genes in each cell type. scITD correctly prioritized the relevant donors and genes involved in each pattern (Figure S1B, S1C, and S1D). We further downsampled the simulated and IFN-beta datasets to assess the impact of cell number on performance, and we observed good performance above an average of 60 cells per donor per cell type (Figure S1E and S1G). We also developed an approach to determine the appropriate number of factors into which the initial tensor should be decomposed (Methods). When applying this approach to the simulated dataset, it recommended a dimensionality that yielded accurate recovery (high AUCs) of true DE genes (Figure S1F and S1D).

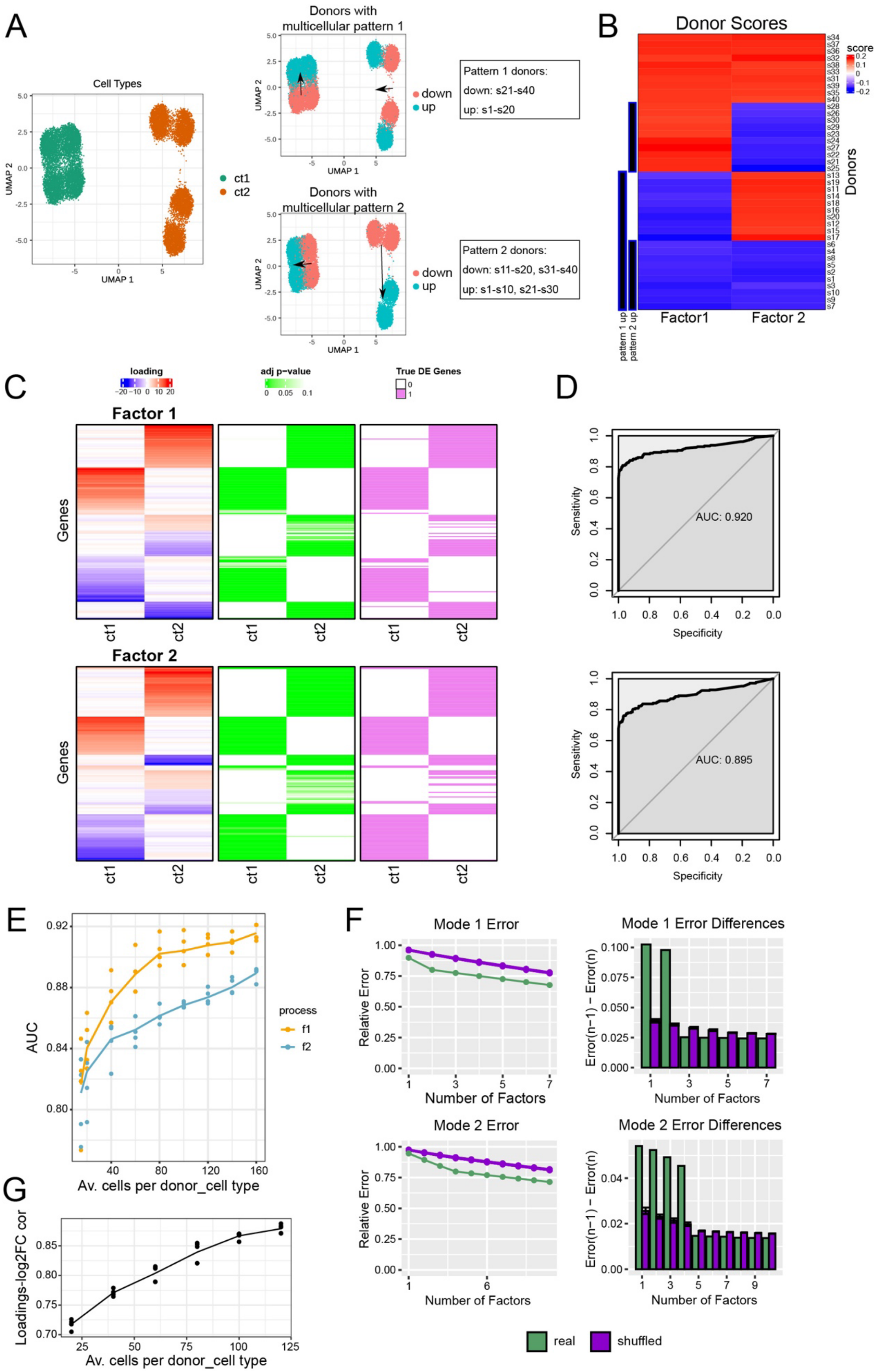

A
Cell Types
UMAP 1
UMAP 2
ct1
ct2
Donors with multicellular pattern 1
down
up
Pattern 1 donors:
down: s21-s40
up: s1-s20
Donors with multicellular pattern 2
down
up
Pattern 2 donors:
down: s11-s20, s31-s40
up: s1-s10, s21-s30
B
Donor Scores
score
0.2
0.1
0
−0.1
−0.2
s34
s37
s36
s32
s38
s33
s31
s39
s35
s40
s28
s26
s30
s29
s23
s24
s27
s22
s21
s25
s13
s19
s11
s14
s18
s16
s20
s12
s15
s17
s6
s4
s8
s5
s2
s1
s3
s10
s9
s7
Donors
pattern 1 up
pattern 2 up
Factor1
Factor 2
C
loading
−20 −10 0 10 20
adj p-value
0 0.05 0.1
True DE Genes
0
1
Factor 1
Genes
ct1
ct2
Factor 2
D
Sensitivity
Specificity
AUC: 0.920
AUC: 0.895
1.0
0.8
0.6
0.4
0.2
0.0
E
AUC
0.92
0.88
0.84
0.80
40
80
120
160
Av. cells per donor_cell type
process
f1
f2
F
Mode 1 Error
Mode 1 Error Differences
Mode 2 Error
Mode 2 Error Differences
Relative Error
Error(n−1) − Error(n)
Number of Factors
real
shuffled
G
Loadings-log2FC cor
0.85
0.80
0.75
0.70
25
50
75
100
125
Av. cells per donor_cell type

Figure S1. Evaluating the performance of scITD with simulated data.

(A) UMAP of the simulated scRNA-seq dataset with two cell types (left). Also shown are two multicellular patterns that separate groups of donors (right). Arrows point to the cells from donors with upregulation of a given multicellular pattern.

(B) Donor scores matrix after applying scITD to extract two factors. Rows are hierarchically clustered. Labels indicating which donors were assigned to have upregulation of each multicellular pattern are shown on the left annotation.

(C) Loadings matrices for the two factors limited to significant genes only. The left matrices show the loading values, the middle matrices show the association significance p-values of each gene in each cell type with the factor, and the right matrices show the true DE genes in each cell type (as set with simulation parameters). Rows are hierarchically clustered.

(D) ROC curves and AUC values for predicting ground truth DE genes in each cell type from each gene's expression-factor association p-value.

(E) Performance of the method on the simulated dataset downsampled to a varying average number of cells per donor per cell type. AUC is computed the same way as in (D) and is shown for each multicellular pattern as distinct colors. Each point represents a different downsampling iteration (n=5), and the line is the mean AUC at a given dataset size.

(F) Rank determination by SVD applied to the full simulated dataset. The left plots show the relative error when performing SVD on the unfolded tensor to varying numbers of factors. Mode 1 error refers to the reconstruction error when SVD is run on the tensor unfolded along the donor dimension. Mode 2 error refers to the reconstruction error when the SVD is run on the tensor unfolded along the gene dimension. The right plots show the change in relative error when incrementing the number of factors. Green bars show the results for the full simulated dataset, whereas purple bars show the results for the dataset after randomly shuffling cell-to-donor assignments (n=50 shuffling iterations). Error bars for the shuffled samples represent standard deviation.

(G) Spearman correlations between loadings and log2FC from the IFN-beta experiment data downsampled to a varying average number of cells per donor per cell type. Each point represents a different downsampling iteration (n=5), and the line is the mean Spearman correlation at a given dataset size.

Analysis of an SLE dataset identifies novel multicellular patterns that stratify patients

Next, we applied scITD to a large scRNA-seq dataset of PBMCs from 115 SLE patients and 56 healthy donors (Figure 2A). We focused our analysis on 7 cell types annotated at a coarse-grained level so that we would have a sufficient number of cells per donor per cell type (Figure 2B left). After transforming expression counts to pseudobulked counts, we further applied batch correction, as groups of donors were pooled and processed together in different 10X Chromium lanes (Methods). We then applied scITD to extract 7 factors, several of which were significantly associated with metadata variables such as SLE status, sex, ethnicity, and age (Figure 2C).

We first investigated factor 1, which had a strong association with SLE status and explained the most variation in the dataset (Figure 2C). The loadings matrix for this factor (Figure 2D top) revealed a core expression program consisting of interferon-stimulated genes (ISGs) impacting multiple cell types. Since the ISGs had large positive loadings, the interpretation is that they are upregulated in donors with large positive sample scores (Figure 2D top and 2E). Therefore, this factor distinguishes SLE patients from healthy donors and delineates those patients who have high ISG expression across all cell types. Interferon signaling has been reported by many other groups in the context of SLE, and it is often present in roughly half of all SLE patients (Baechler et al., 2003; Bennett et al., 2003; Crow et al., 2003; Han et al., 2003; Hooks et al., 2010; Nehar-Belaid et al., 2020).

Applying gene set enrichment analysis (GSEA) per cell type for this factor (Figure 2D bottom) yielded enrichment of the “response to type I interferon” gene set in all cell types as well as other gene sets enriched in specific cell types, especially monocytes. Some of the monocyte-specific gene sets included interleukin (IL)-1 production, IL-6 production, IL-10 production, and TNF production among others (Figure 2D bottom). These may simply represent the monocyte-specific responses to interferon. Supporting this, roughly 70% of the significant genes in cMonocytes were differentially expressed for this cell type in the IFN-beta stimulation experiment discussed above (adjusted p-values < 0.05). The ISG-high donors also showed higher expression of genes from various other biological processes including cell-cycle in CD8+ T cells (Tc), apoptosis in NK and Tc cells, and proteolysis in multiple cell types (Figure 2D bottom). Certain proteasome subunits (*PSMB8*, *PSMB9*, and *PSMB10*) are reported to be upregulated by interferon during infections, functioning to enhance antigen presentation (Yang et al., 1992; Shin et al., 2006). Interestingly, these SLE patients also had upregulation of a pathway for regulatory T cell differentiation (Figure 2D bottom) marked by the elevated expression of the canonical Treg transcription factor, *FOXP3*, among helper T cells in these donors (Figure 2D top). Consistent with this, we observed a significant increase in Treg cell proportions (Tregs are a subcluster within Th cells) for donors with high factor 1 scores (p-value=$8.7x10^{-19}$). Previous studies have also shown increased numbers of Tregs in SLE patients compared to healthy donors and often accompanying high ISG expression (Suen and Chiang, 2012; Ferreira et al., 2019). In addition to this finding, we also found a significant reduction in the proportion of naïve Th cells for donors with high ISG expression (p-value = 0.001). Whereas the naïve Th association was also shown in the main analysis of this dataset (Perez et al.), the Treg association was not previously described. This highlights the ability of scITD to identify coordinated cell-type-specific transcriptional modules within a cohort of patients with autoimmunity.

When using scITD on the data without any batch correction, we were able to isolate batch effects into distinct factors (Figure 2F). Notably, the number of batch-associated factors remained stable when continually increasing the total number of factors. Therefore, a user can apply scITD to non-batch-corrected data and then choose to analyze the factors that are not batch-associated. However, this also provides a unique opportunity to explicitly study the multicellular nature of batch effects. We demonstrated this using our SLE dataset and showed that 10X Chromium lane-associated factors often have consistent expression patterns across different cell types (Figure S2A and S2B). We further analyzed these patterns for associations with gene-level attributes such as GC-content, ambient RNA expression, or gene ontologies. For example, 10X lane-associated genes shared between cell types were more likely to be present at higher fractions in droplets containing ambient RNA compared to the cell-type-specific genes (Figure S2C). This analysis provides a unique use case for scITD to examine how batch effects impact multiple cell types and may eventually pave the way for improved batch-correction techniques.

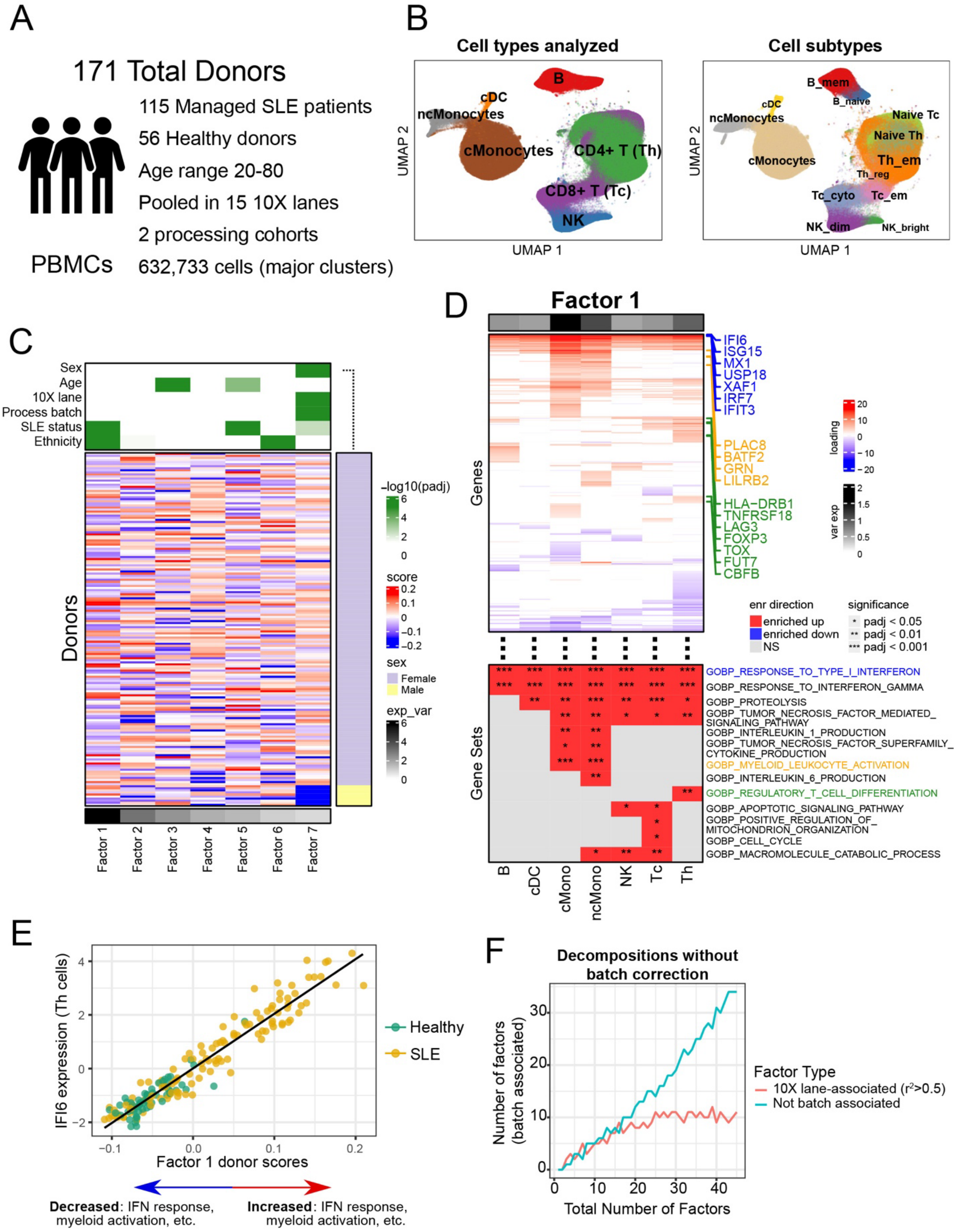

A
171 Total Donors
115 Managed SLE patients
56 Healthy donors
Age range 20-80
Pooled in 15 10X lanes
2 processing cohorts
PBMCs
632,733 cells (major clusters)
B
Cell types analyzed
B
cDC
ncMonocytes
cMonocytes
CD4+ T (Th)
CD8+ T (Tc)
NK
UMAP 2
UMAP 1
Cell subtypes
B_mem
B_naive
cDC
ncMonocytes
cMonocytes
Naive Tc
Naive Th
Th_em
Th_reg
Tc_cyto
Tc_em
NK_dim
NK_bright
UMAP 2
UMAP 1
C
Sex
Age
10X lane
Process batch
SLE status
Ethnicity
Donors
Factor 1
Factor 2
Factor 3
Factor 4
Factor 5
Factor 6
Factor 7
-log10(padj)
6
4
2
0
score
0.2
0.1
0
−0.1
−0.2
sex
Female
Male
exp_var
6
4
2
0
D
Factor 1
Genes
IFI6
ISG15
MX1
USP18
XAF1
IRF7
IFIT3
PLAC8
BATF2
GRN
LILRB2
HLA-DRB1
TNFRSF18
LAG3
FOXP3
TOX
FUT7
CBFB
loading
20
10
0
− 10
− 20
var exp
2
1.5
1
0.5
0
enr direction
enriched up
enriched down
NS
significance
* padj < 0.05
** padj < 0.01
*** padj < 0.001
Gene Sets
GOBP_RESPONSE_TO_TYPE_I_INTERFERON
GOBP_RESPONSE_TO_INTERFERON_GAMMA
GOBP_PROTEOLYSIS
GOBP_TUMOR_NECROSIS_FACTOR_MEDIATED_SIGNALING_PATHWAY
GOBP_INTERLEUKIN_1_PRODUCTION
GOBP_TUMOR_NECROSIS_FACTOR_SUPERFAMILY_CYTOKINE_PRODUCTION
GOBP_MYELOID_LEUKOCYTE_ACTIVATION
GOBP_INTERLEUKIN_6_PRODUCTION
GOBP_REGULATORY_T_CELL_DIFFERENTIATION
GOBP_APOPTOTIC_SIGNALING_PATHWAY
GOBP_POSITIVE_REGULATION_OF_MITOCHONDRION_ORGANIZATION
GOBP_CELL_CYCLE
GOBP_MACROMOLECULE_CATABOLIC_PROCESS
B
cDC
cMono
ncMono
NK
Tc
Th
E
IFI6 expression (Th cells)
4
2
0
−2
−0.1
0.0
0.1
0.2
Factor 1 donor scores
Healthy
SLE
Decreased: IFN response, myeloid activation, etc.
Increased: IFN response, myeloid activation, etc.
F
Decompositions without batch correction
Number of factors (batch associated)
30
20
10
0
0
10
20
30
40
Total Number of Factors
Factor Type
10X lane-associated (r²>0.5)
Not batch associated

Figure 2. SLE scRNA-seq dataset overview and scITD analysis.

(A) Description of the SLE PBMC dataset.

(B) UMAP of single-cell gene expression from the SLE dataset, showing the coarse clustering used in the downstream analyses (left) and cell subtype annotations (right).

(C) Donor scores heatmap with metadata association p-values annotated at the top. The p-values were calculated using univariate linear model F-tests. Rows are grouped by the sex of each donor, and this is shown on the right of the heatmap. Columns are ordered by explained variance for each factor, and this is displayed at the bottom of the heatmap.

(D) Loadings matrix for factor 1 limited to only significant genes (top) with select GSEA enriched gene sets in each cell type (bottom). GSEA p-values were calculated using the FGSEA R package with Gene Ontology Biological Process (GOBP) gene sets. The top annotation shows the percent of overall explained variance for each cell type of the factor. The genes highlighted with different colors are a few leading-edge genes for the gene sets with corresponding colors. The rows of both heatmaps are hierarchically clustered.

(E) Expression of the ISG, *IFI6*, in Th cells plotted against donor scores for factor 1. Each point is a donor. Points are colored by the donor's SLE status. The p-value was calculated using a linear model F-test. Arrows at the bottom highlight a few of the other biological processes that co-occur with ISG expression.

(F) The number of batch-associated factors ($r^2 > 0.5$) at a given total number of factors extracted from the SLE dataset when no batch correction is applied.

One unique attribute of the Tucker tensor decomposition is that the factors can be rotated to improve their interpretability (Unkel et al., 2011; Zhou and Cichocki, 2012). Therefore, we compared rotations applied to either the loadings or the donor scores of the SLE decomposition (Figure S3A) (Methods). We show that by rotating the loadings, the factor patterns become less similar to one another and more modular (Figure S3B). This independence enables us to interpret individual factors as functionally interconnected multicellular patterns. In contrast, a rotation of the donor scores can produce factors that more strongly stratify patients into groups, although the multicellular patterns become less modular (Figure S3C and S3B). From a user perspective, it is important to keep these distinctions in mind and to select an appropriate rotation based on the goals of the analysis. For most analyses in this study, we opted to use a rotation on loadings.

In addition to identifying multicellular patterns broadly associated with disease status, scITD can also be used to identify those that stratify samples, which may be particularly useful for studying a heterogeneous disease like SLE. Therefore, we next applied scITD to only the samples from SLE patients to help identify multicellular patterns that may be associated with various clinical annotations. The resulting decomposition yielded factors that were highly similar to the previous ones (Figure S4C). Factors 1, 2, 3, and 7 involve multiple cell types, factors 4 and 5 primarily involve cytotoxic T cells (Tc) and Th cells, and factor 6 shows consistent expression variation for all cell types, as this factor is associated with sex (Figure S4A and S4B). We further demonstrated that the factors have high stability by randomly subsampling to 85% of the donors, recomputing the decomposition, and assessing factor correlations with those from the full SLE-only decomposition (Figure S4D).

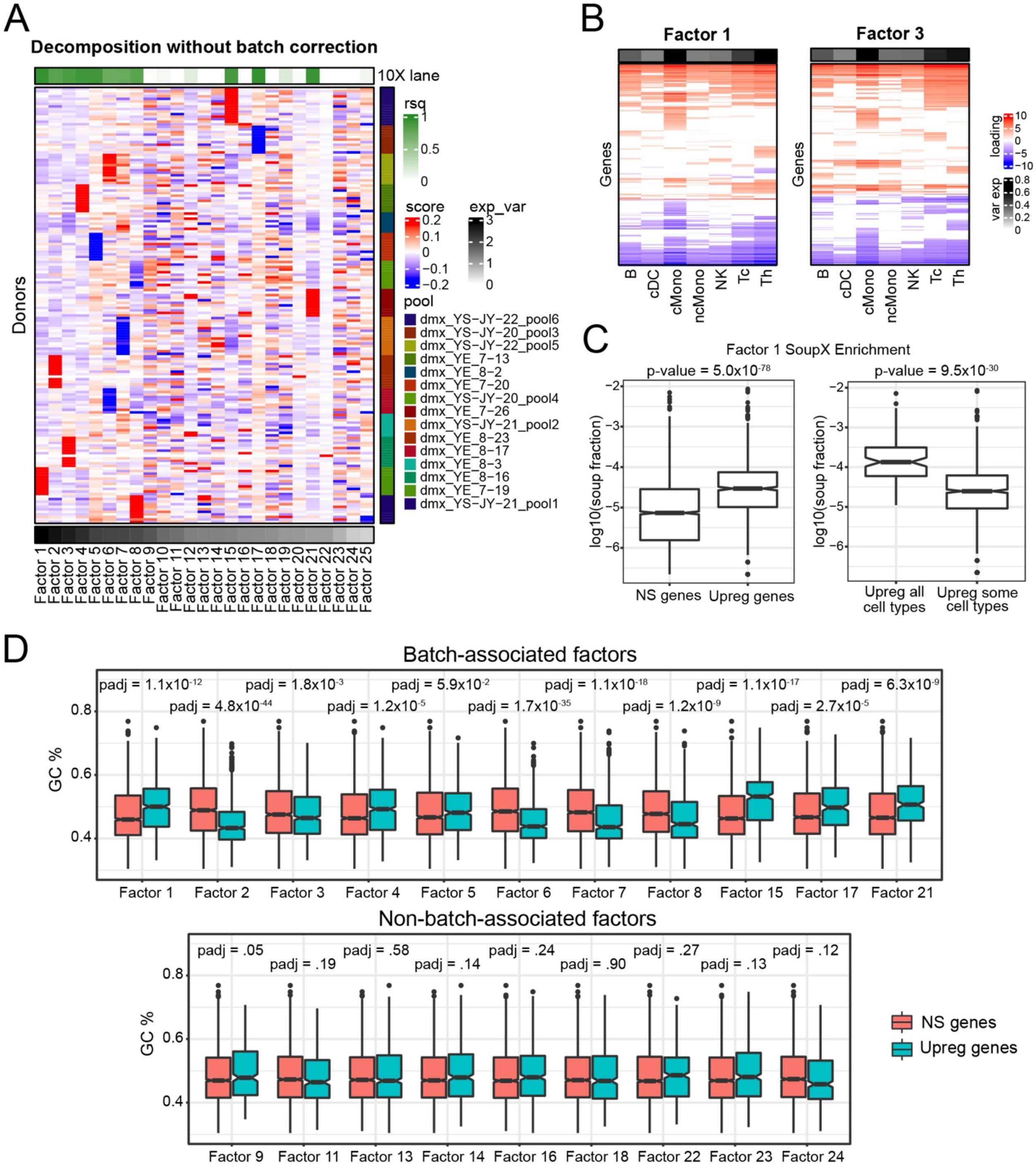


Figure S2. Analysis of technical effects through extracted batch-associated factors.

(A) Donor scores matrix for a decomposition of the SLE dataset without applying batch-effect removal. Also shown are p-values for associations between the factor scores and 10X lane (top). The p-values were calculated using univariate linear model F-tests. Columns are ordered by explained variance, shown as a bottom annotation. Rows are grouped by 10X lane, shown as an annotation on the right side.

(B) Loadings matrices for two factors associated with 10X Chromium lanes limited to the significant genes only. Rows are hierarchically clustered.

(C) Analysis of factor 1 loadings for association with ambient RNA content of batch dmx_YE_7-19. The left plot shows the association between all genes that are upregulated in this batch and their fractional representation in the ambient RNA "soup" for the batch. The right plot shows ambient RNA fractions for genes that were significantly upregulated across all cell types compared to those upregulated only in some cell types. Associations are calculated by two-sample t-tests for (G-H). "NS genes" for (G-H) includes all genes that were not significantly upregulated.

(D) Associations between upregulated genes in each factor and GC content. This is shown separately for the batch-associated factors (top) or non-batch-associated factors (bottom).

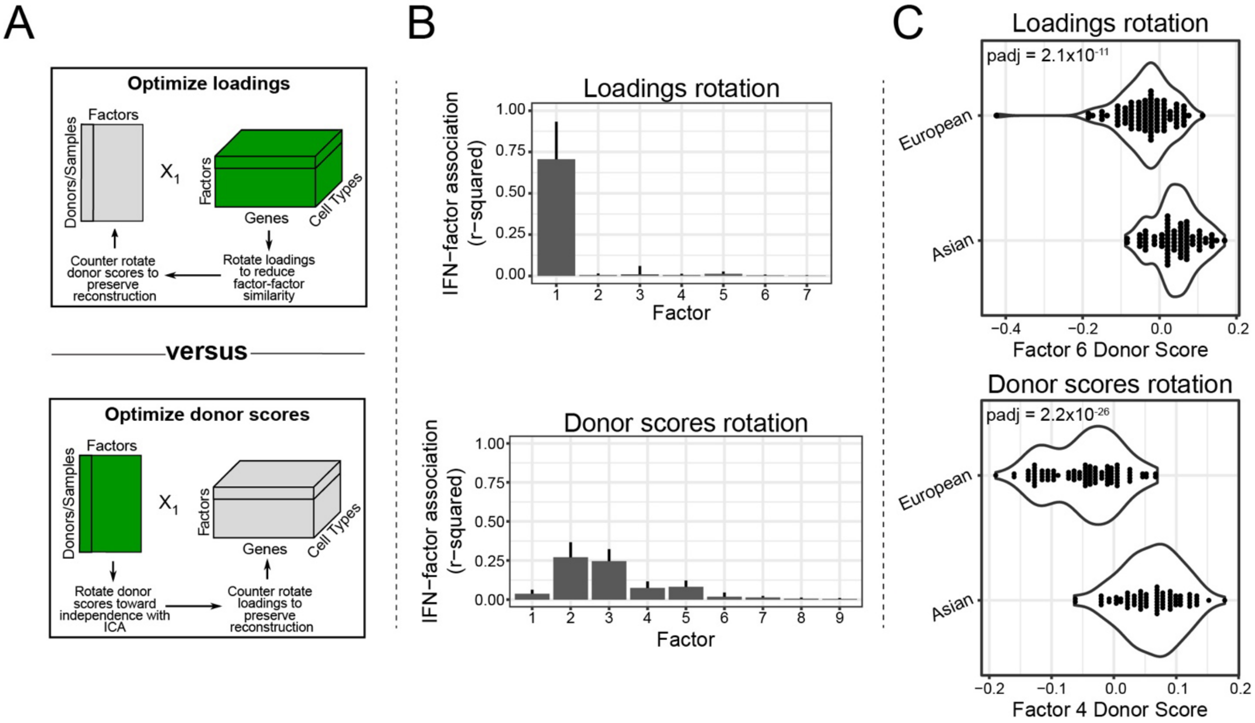


Figure S3. Comparing the impact of different factor rotations.

(A) The two general approaches taken to rotating factors. Either the loadings are rotated to maximize some criterion (top), or the donor scores are rotated to maximize some criterion (bottom). In either case, the non-optimized component of the decomposition output is counter-rotated to preserve the original reconstruction error.

(B) Average ISG-factor score associations (r-squared) for the loadings rotation (top) and donor scores rotation (bottom). The ISGs used are *ISG15, IFI27, IRF7, HERC5, LY6E, MX1, OAS2, OAS3, RSAD2, USP18,* and *GBP5*. A separate r-squared value is calculated for each cell type. The error bars represent the standard deviation across this set of genes.

(C) Comparing the strength of association for the ethnicity-associated factor from the loadings rotation (top) and the donor scores rotation (bottom). The p-values were calculated using univariate linear model F-tests.

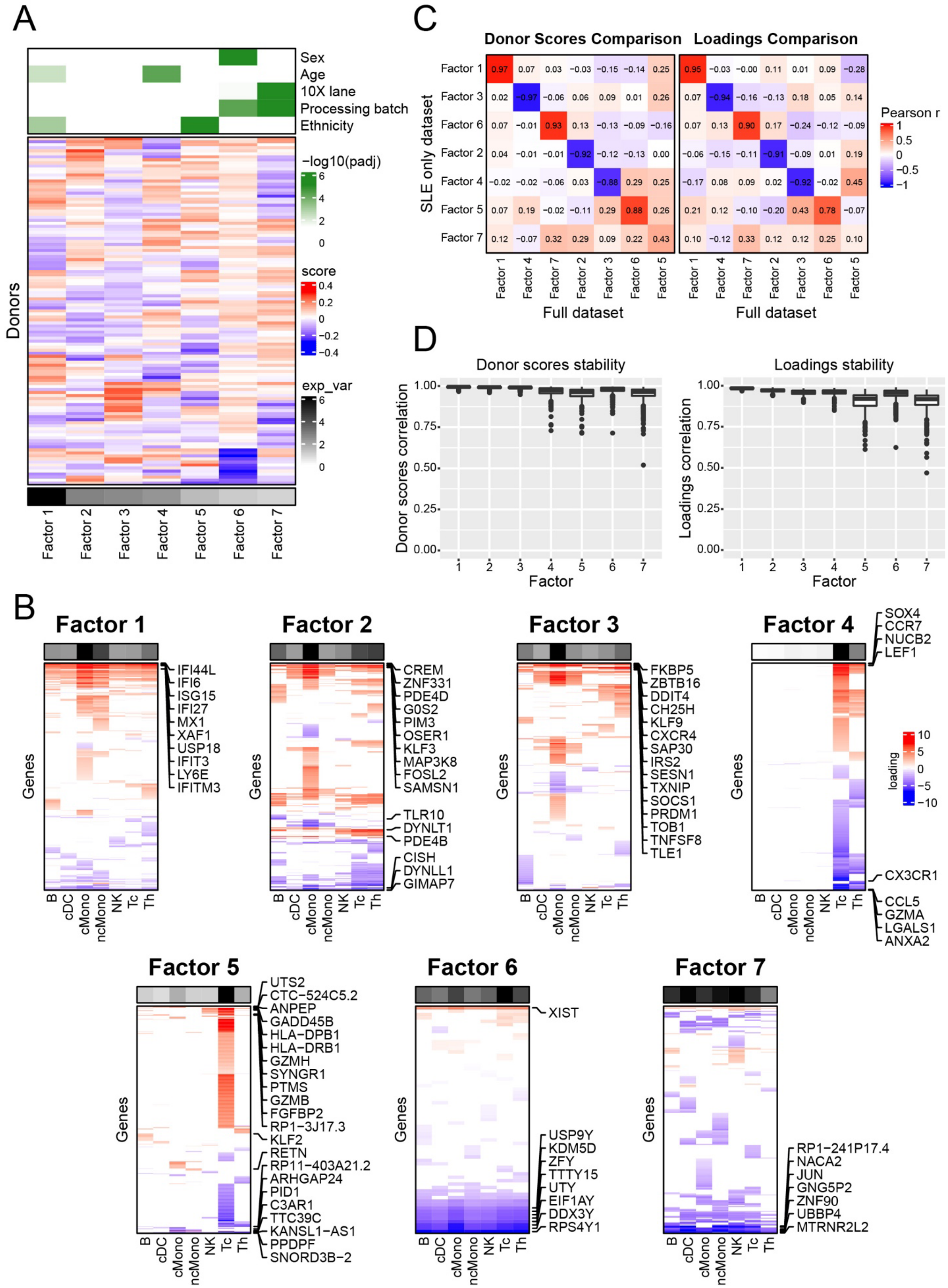

A
Sex
Age
10X lane
Processing batch
Ethnicity
−log10(padj)
6
4
2
0
score
0.4
0.2
0
−0.2
−0.4
exp_var
6
4
2
0
Donors
Factor 1
Factor 2
Factor 3
Factor 4
Factor 5
Factor 6
Factor 7
C
Donor Scores Comparison
Loadings Comparison
SLE only dataset
Factor 1 0.97 0.07 0.03 −0.03 −0.15 −0.14 0.25 | 0.95 −0.03 −0.00 0.11 0.01 0.09 −0.28
Factor 3 0.02 −0.97 −0.06 0.06 0.09 0.01 0.26 | 0.07 −0.94 −0.16 −0.13 0.18 0.05 0.14
Factor 6 0.07 −0.01 0.93 0.13 −0.13 −0.09 −0.16 | 0.07 0.13 0.90 0.17 −0.24 −0.12 −0.09
Factor 2 0.04 −0.01 −0.01 −0.92 −0.12 −0.13 0.00 | −0.06 −0.15 −0.11 −0.91 −0.09 0.01 0.19
Factor 4 −0.02 −0.02 −0.06 0.03 −0.88 0.29 0.25 | −0.17 0.08 0.09 0.02 −0.92 −0.02 0.45
Factor 5 0.07 0.19 −0.02 −0.11 0.29 0.88 0.26 | 0.21 0.12 −0.10 −0.20 0.43 0.78 −0.07
Factor 7 0.12 −0.07 0.32 0.29 0.09 0.22 0.43 | 0.10 −0.12 0.33 0.12 0.12 0.25 0.10
Factor 1 Factor 4 Factor 7 Factor 2 Factor 3 Factor 6 Factor 5
Full dataset
Pearson r
1
0.5
0
−0.5
−1
D
Donor scores stability
Loadings stability
Donor scores correlation
Loadings correlation
1.00
0.75
0.50
0.25
0.00
1 2 3 4 5 6 7
Factor
B
Factor 1
IFI44L
IFI6
ISG15
IFI27
MX1
XAF1
USP18
IFIT3
LY6E
IFITM3
Factor 2
CREM
ZNF331
PDE4D
G0S2
PIM3
OSER1
KLF3
MAP3K8
FOSL2
SAMSN1
TLR10
DYNLT1
PDE4B
CISH
DYNLL1
GIMAP7
Factor 3
FKBP5
ZBTB16
DDIT4
CH25H
KLF9
CXCR4
SAP30
IRS2
SESN1
TXNIP
SOCS1
PRDM1
TOB1
TNFSF8
TLE1
Factor 4
SOX4
CCR7
NUCB2
LEF1
CX3CR1
CCL5
GZMA
LGALS1
ANXA2
loading
10
5
0
−5
−10
Factor 5
UTS2
CTC−524C5.2
ANPEP
GADD45B
HLA−DPB1
HLA−DRB1
GZMH
SYNGR1
PTMS
GZMB
FGFBP2
RP1−3J17.3
KLF2
RETN
RP11−403A21.2
ARHGAP24
PID1
C3AR1
TTC39C
KANSL1−AS1
PPDPF
SNORD3B−2
Factor 6
XIST
USP9Y
KDM5D
ZFY
TTTY15
UTY
EIF1AY
DDX3Y
RPS4Y1
Factor 7
RP1−241P17.4
NACA2
JUN
GNG5P2
ZNF90
UBBP4
MTRNR2L2
Genes
B
cDC
cMono
ncMono
NK
Tc
Th

Figure S4. Applying scITD to the SLE-only portion of the dataset.

(A) Donor scores heatmap with metadata association p-values annotated at the top. The p-values are calculated using univariate linear model F-tests. Columns are ordered by explained variance for each factor, which is shown at the bottom. Rows are hierarchically clustered.

(B) Loadings matrices for all factors of the decomposition, reduced to only the significant genes in each cell type. Top annotations represent explained variance for each cell type of the factor. Rows are hierarchically clustered. Genes that have the strongest associations with each factor are shown as labeled callouts.

(C) Pearson correlations between factors from the SLE-only dataset and factors from the full dataset decomposition. The left heatmap shows correlations between factor donor scores of the two decompositions. The right heatmap shows correlations between factor loadings for the two decompositions. Rows and columns are hierarchically clustered.

(D) Stability analysis results for the SLE-only decomposition over 500 iterations. In each iteration, the dataset was subsampled to 85% of the donors. Values represent the maximum absolute value factor-factor correlation coefficients for each of the original factors mapped to the factors from each subsampled dataset decomposition. This is shown separately for max donor score correlations (left) and max loadings correlations (right).

Next, we associated these seven factors (from the SLE-only decomposition) with 41 clinical features including SLE symptoms and medication use. One of these is the SLE disease activity index (SLEDAI), which is a composite score derived from 24 manifestations and immunologic features spanning 9 organ systems (Bombardier et al., 1992). Factor 1, which was again described by pan-cellular ISG expression, was significantly associated with the presence of anti-dsDNA and anti-smith autoantibodies (Figure 3A) as well as higher SLEDAI scores (Figure 3B). The connection between higher ISG expression and these clinical features has been previously shown in several studies (Hooks et al., 1979; Bennett et al., 2003; Kirou et al., 2005; Nikpour et al., 2008; Weckerle et al., 2011). The Pearson correlation between factor 1 sample scores and SLEDAI was 0.321, which is similar to that which was observed in other studies (Catalina et al., 2019; Enocsson et al., 2021; Juárez-Vicuña et al., 2021).

Lupus nephritis is one of the most severe complications of SLE, and anti-dsDNA autoantibodies are a critical though insufficient component to its development (Yung and Chan, 2015). Therefore, we sought to identify multicellular patterns that are associated with nephritis when autoantibodies are present. Factor 2 exhibited a significant association with the frequency of lupus nephritis among SLE patients who were positive for anti-dsDNA autoantibodies (Figure 3C). The frequency of patients with nephritis was computed using a sliding window along the factor scores to calculate the number of patients positive for lupus nephritis among those positive for anti-dsDNA autoantibodies (Methods). Notably, there is no association in patients negative for anti-dsDNA autoantibodies, and co-occurrence of anti-smith autoantibodies with lupus nephritis for this factor was less significant (p-value ~ 0.05). This finding suggests that the factor 2 multicellular pattern might exacerbate the adverse effects of anti-dsDNA autoantibodies specifically, leading to lupus nephritis. GSEA analysis reveals that factor 2 is enriched for stress response as well as altered cell cycle, apoptosis, cell migration, and cell adhesion pathways in multiple cell types (Figure 3D). Directionally, the genes in these pathways were overexpressed in the patients with high factor 2 scores. The observation of p38 MAPK signaling in several cell types was also intriguing because there is evidence that this pathway is an important mediator in the development of lupus nephritis, acting in concert with anti-dsDNA autoantibodies (Iwata et al., 2003; Yung et al., 2010; Liu et al., 2016).

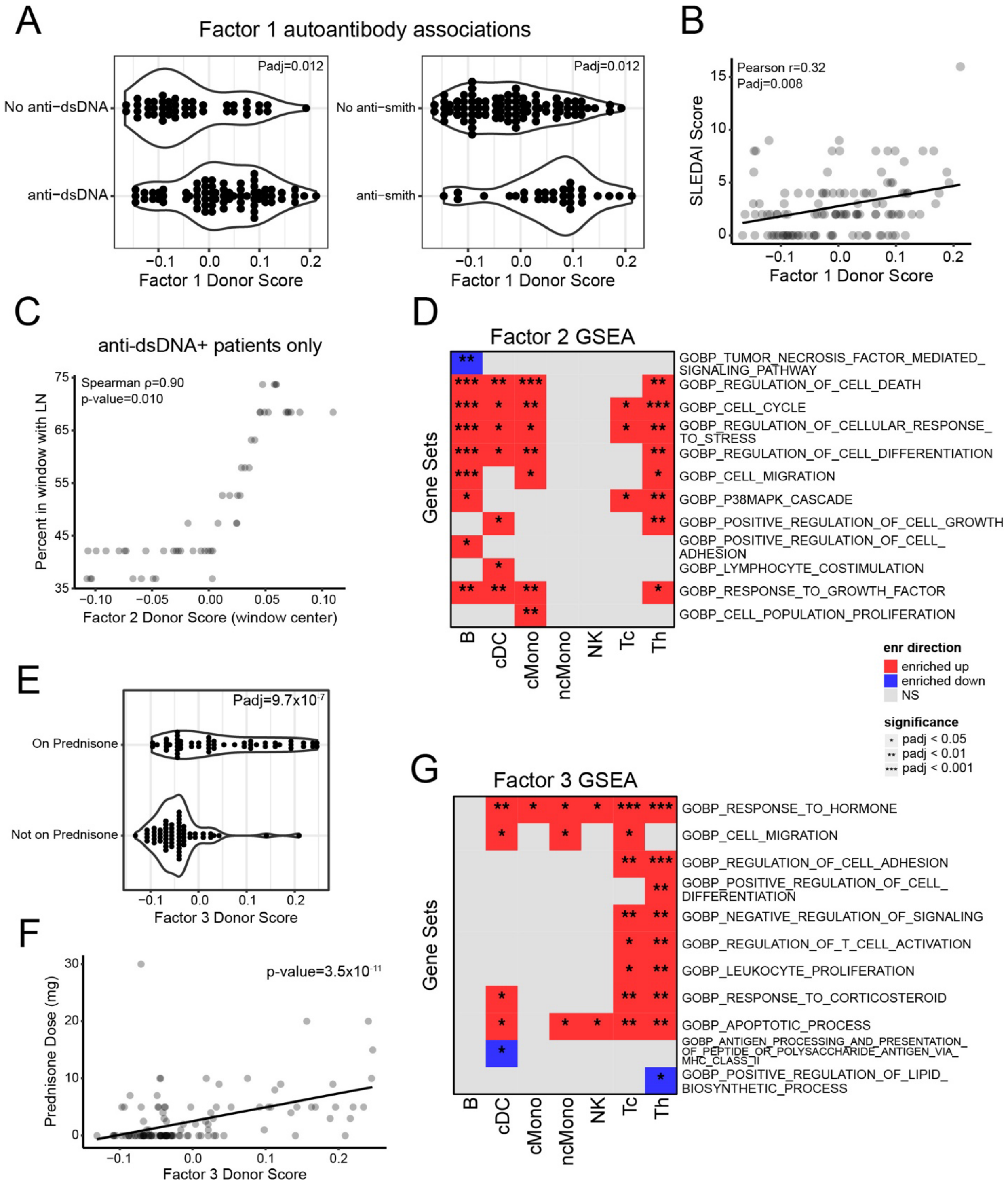


Figure 3. SLE decomposition factor associations with clinical covariates.

(A) Factor 1 associations with anti-dsDNA autoantibody presence (left) and anti-smith autoantibody presence (right). The significance of each association was computed using logistic regression with a likelihood-ratio test.

(B) Factor 1 association with SLEDAI score. Statistical significance was calculated using an ordinal logistic regression (Methods). The line is a linear model.

(C) Factor 2 association with frequency of lupus nephritis among patients positive for anti-dsDNA autoantibodies. A sliding window was used to compute the percent of patients within the window that had lupus nephritis (Methods). Each point represents the sliding window center.
(D) Factor 2 select enriched gene sets computed for each cell type. Enrichment significance was calculated using the FGSEA R package with GOBP gene sets (also applies to G). Rows are hierarchically clustered.
(E) Factor 3 association with prednisone use. The significance of each association was computed using logistic regression with a likelihood-ratio test.
(F) Factor 3 association with prednisone dose. The outlier with the highest prednisone dose was not included in the calculation of the linear model p-value but is still shown in the plot. The p-value was calculated using a linear model F-test. The line is a linear model.
(G) Factor 3 select enriched gene sets computed for each cell type. Enrichment significance was calculated using the FGSEA R package with GOBP gene sets. Rows are hierarchically clustered.

Association analysis of treatment revealed that factor 3 was strongly associated with both the use and dosage of the corticosteroid prednisone (Figure 3E and 3F). GSEA analysis for this factor confirmed the expected enrichment of corticosteroid and hormone response genes in multiple cell types (Figure 3G). There were also several enriched pathways specific to Th cells. We further noted that several of the SLE patients taking prednisone did not have high scores for this factor. In fact, the SLE patient taking the highest prednisone dose had practically none of this multicellular pattern (Figure 3F). These few individuals may exhibit resistance to prednisone, as previous reports have shown that up to a third of SLE patients may have some degree of resistance to the medication (Luijten et al., 2013). Overall, this illustrates one of the main benefits of using scITD over simple DE analyses, as DE would be underpowered to detect the prednisone-associated genes with such outliers.

Inference of ligand-receptor (LR) interactions reveals potential mediators of multicellular patterns

We also sought to identify LR interactions that are candidate mediators of the multicellular patterns identified by scITD. Our approach to inferring LR interactions differs from standard approaches in several ways. Most single-cell LR methods identify interactions based on the upregulation of ligands and their cognate receptors in pairs of cell clusters without regard to the sample of origin. Here, we explicitly test for interactions that are differentially active between samples. This is done by associating ligand expression in a source cell type with the expression of WGCNA co-expression gene modules in a target cell type, which we use as a proxy to represent the downstream effects of an LR signaling event (Figure 4A) (Methods). Applying this strategy to the SLE samples, we identified a set of ligands and gene modules significantly associated with each other across samples (Figure 4B left). These candidate interactions can be further associated with the inferred multicellular patterns (Figure 4B right). We further demonstrated that our approach enriches for plausible LR interactions and outperforms a standard LR inference approach (Methods) (Figure S5A and S5B).

Next, we more closely examined several of the top candidate LR interactions. One such prediction is the interaction of the ligand *TNFSF13B* expressed from cMonocytes with its cognate receptor on B cells. Specifically, we identified a positive association between monocytic expression of the ligand and a co-expression gene module in the B cells, B_m1 (Figure 4B blue arrows and 4C left). The protein that *TNFSF13B* codes for is more commonly referred to as the B-lymphocyte stimulator (BLyS) or B-cell activating factor (BAFF). Notably, this protein is the target of the recently developed therapeutic, belimumab, which acts to reduce B cell activation and autoantibody production in SLE (Allen et al., 2021). Here, the genes in the associated B_m1 co-expression module were significantly enriched for genes in B-cell receptor activation gene sets as well as differentiation and cell-cycle gene sets (adjusted p-values < 0.05), as would be expected from BAFF stimulation. We observed that donors with high ISG expression (high factor 1 scoring donors) had significantly higher expression of *TNFSF13B* (Figure 4B right and 4C right). This is consistent with previous reports that the *TNFSF13B* ligand is itself a known target upregulated by interferon (Sjöstrand et al., 2016). As we have shown, connecting our LR inference results to our scITD multicellular patterns enables contextualization of candidate interactions, improving our confidence in them.

Another particularly strong candidate that we identified is the ICOSLG-ICOS interaction from cMonocytes to Th cells (Figure 4D). Specifically, we found that donors with higher *ICOSLG* expression in cMonocytes had significantly higher expression of gene module Th_m5 in Th cells. The Th_m5 gene module was enriched for genes involved in T-cell receptor activation, cell cycle, and p38 MAPK pathways (Figure 4E) consistent with the known co-stimulatory role for ICOSLG-ICOS binding in T cell activation (Dodeller and Schulze-Koops, 2006; Wikenheiser and Stumhofer, 2016). We also observed that genes in the Th_m5 module had significantly higher NicheNet regulatory potential scores compared to genes in other modules (Figure 4F). For a given ligand, genes with higher regulatory potential scores are more likely to be downstream targets of that ligand (Browaeys et al., 2020). Therefore, this result is consistent with the hypothesis that donors who overexpress *ICOSLG* have increased functional ICOSLG-ICOS signaling between monocytes and Th cells. We also noted that the donors with increased *ICOSLG* expression had significantly higher scores for factor 2 (Figure 4B right and 4D right), suggesting the interaction between ICOSLG-ICOS may promoting T cell proliferation and MAPK activation to increase the frequency of lupus nephritis (Figure 4G). Corroborating this result, a previous study showed that T cell ICOS stimulation by myeloid cells contributed to the development of lupus nephritis and was mediated by increased T cell survival (Teichmann et al., 2015).

We also identified a significant interaction channel that appeared active for donors with functional prednisone response. Specifically, we found that the ligand *THBS1* in cDCs was positively correlated with the Th_m9 co-expression module (Figure 4H left). The *THBS1* ligand itself was also positively correlated with factor 3 donor scores, indicating that it is overexpressed with functional prednisone response (Figure 4H right). Similar to factor 3 (Figure 3G), module Th_m9 was significantly enriched (adjusted p-values < 0.01) for genes involved in response to hormone (e.g., *KLF9*, *TXNIP*) and regulation of T cell activation (e.g., SOCS1, *NFKBIZ*). Previous studies have also shown that dendritic cell-derived *THBS1* can promote Treg development when interacting with integrin-associated protein (CD47) (Grimbert et al., 2006). Therefore, we tested for an association between the factor 3 scores and Treg proportions. This association was statistically significant (p-value = 0.002), with a relative expansion of this Th subpopulation in the prednisone patients. Unlike the factor 1-high donors, however, the factor 3-high patients did not have increased ISG expression. We also observed significantly higher *THBS1* regulatory potential scores for genes in the Th_m9 gene module compared to genes in all other modules (p-value = 2.8×10^{-8}), supporting the inference of this interaction. These results highlight a high confidence shift in intercellular communication upon prednisone use that may mechanistically contribute to its anti-inflammatory effects.

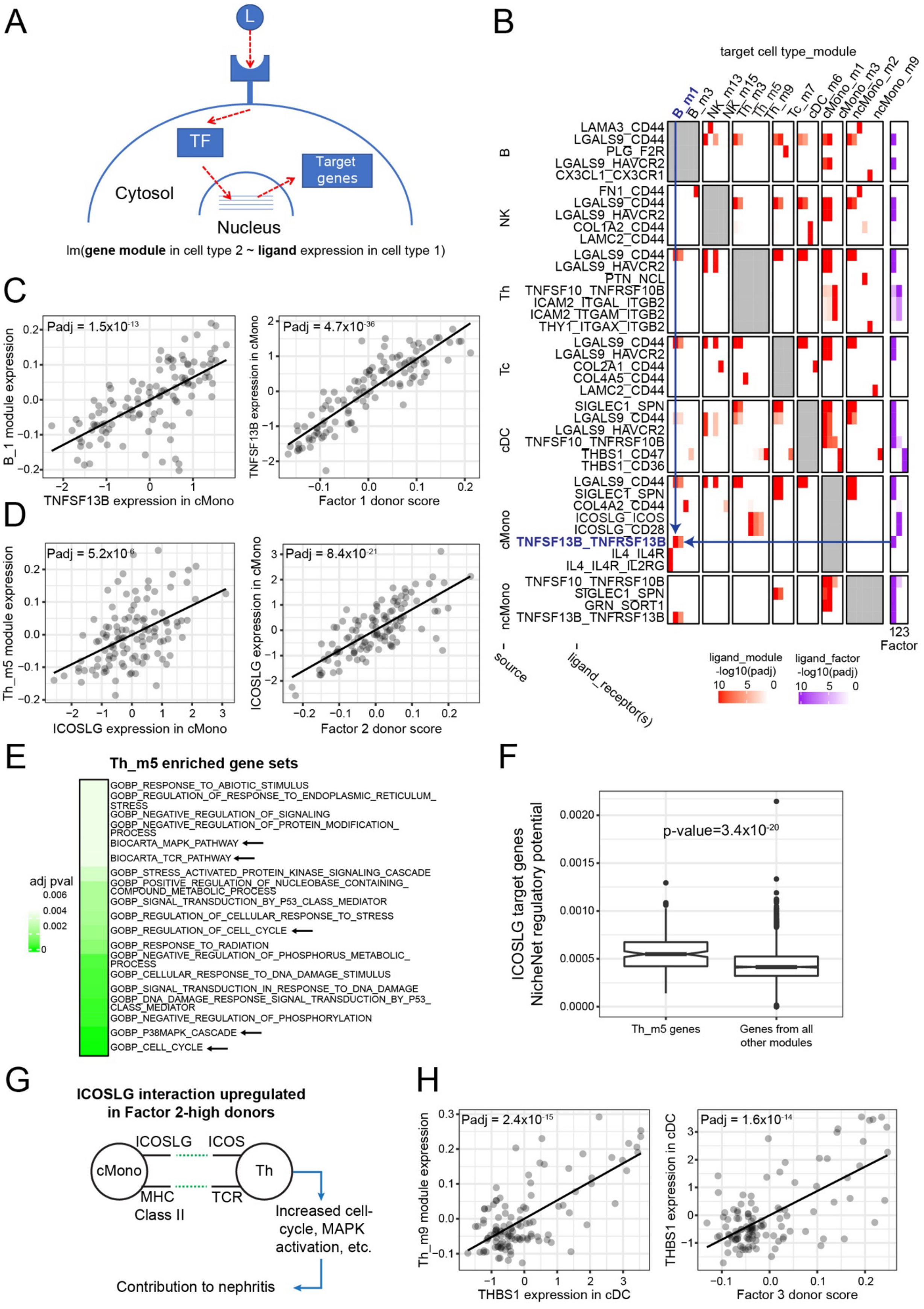

A
L
TF
Target genes
Cytosol
Nucleus
lm(gene module in cell type 2 ~ ligand expression in cell type 1)
B
target cell type_module
B_m1
B_m3
NK_m13
NK_m15
Th_m3
Th_m5
Th_m9
Tc_m7
cDC_m6
cMono_m1
cMono_m3
ncMono_m2
ncMono_m9
B
LAMA3_CD44
LGALS9_CD44
PLG_F2R
LGALS9_HAVCR2
CX3CL1_CX3CR1
NK
FN1_CD44
LGALS9_CD44
LGALS9_HAVCR2
COL1A2_CD44
LAMC2_CD44
Th
LGALS9_CD44
LGALS9_HAVCR2
PTN_NCL
TNFSF10_TNFRSF10B
ICAM2_ITGAL_ITGB2
ICAM2_ITGAM_ITGB2
THY1_ITGAX_ITGB2
Tc
LGALS9_CD44
LGALS9_HAVCR2
COL2A1_CD44
COL4A5_CD44
LAMC2_CD44
cDC
SIGLEC1_SPN
LGALS9_CD44
LGALS9_HAVCR2
TNFSF10_TNFRSF10B
THBS1_CD47
THBS1_CD36
cMono
LGALS9_CD44
SIGLEC1_SPN
COL4A2_CD44
ICOSLG_ICOS
ICOSLG_CD28
TNFSF13B_TNFRSF13B
IL4_IL4R
IL4_IL4R_IL2RG
ncMono
TNFSF10_TNFRSF10B
SIGLEC1_SPN
GRN_SORT1
TNFSF13B_TNFRSF13B
123
Factor
source
ligand_receptor(s)
ligand_module -log10(padj)
10 5 0
ligand_factor -log10(padj)
10 5 0
C
Padj = 1.5x10-13
B_1 module expression
TNFSF13B expression in cMono
0.2 0.1 0.0 −0.1 −0.2
−2 −1 0 1
Padj = 4.7x10-36
TNFSF13B expression in cMono
Factor 1 donor score
2 1 0 −1 −2
−0.1 0.0 0.1 0.2
D
Padj = 5.2x10-6
Th_m5 module expression
ICOSLG expression in cMono
0.2 0.1 0.0 −0.1 −0.2
−2 −1 0 1 2 3
Padj = 8.4x10-21
ICOSLG expression in cMono
Factor 2 donor score
3 2 1 0 −1 −2
−0.2 −0.1 0.0 0.1 0.2
E
Th_m5 enriched gene sets
adj pval
0.006
0.004
0.002
0
GOBP_RESPONSE_TO_ABIOTIC_STIMULUS
GOBP_REGULATION_OF_RESPONSE_TO_ENDOPLASMIC_RETICULUM_STRESS
GOBP_NEGATIVE_REGULATION_OF_SIGNALING
GOBP_NEGATIVE_REGULATION_OF_PROTEIN_MODIFICATION_PROCESS
BIOCARTA_MAPK_PATHWAY
BIOCARTA_TCR_PATHWAY
GOBP_STRESS_ACTIVATED_PROTEIN_KINASE_SIGNALING_CASCADE
GOBP_POSITIVE_REGULATION_OF_NUCLEOBASE_CONTAINING_COMPOUND_METABOLIC_PROCESS
GOBP_SIGNAL_TRANSDUCTION_BY_P53_CLASS_MEDIATOR
GOBP_REGULATION_OF_CELLULAR_RESPONSE_TO_STRESS
GOBP_REGULATION_OF_CELL_CYCLE
GOBP_RESPONSE_TO_RADIATION
GOBP_NEGATIVE_REGULATION_OF_PHOSPHORUS_METABOLIC_PROCESS
GOBP_CELLULAR_RESPONSE_TO_DNA_DAMAGE_STIMULUS
GOBP_SIGNAL_TRANSDUCTION_IN_RESPONSE_TO_DNA_DAMAGE
GOBP_DNA_DAMAGE_RESPONSE_SIGNAL_TRANSDUCTION_BY_P53_CLASS_MEDIATOR
GOBP_NEGATIVE_REGULATION_OF_PHOSPHORYLATION
GOBP_P38MAPK_CASCADE
GOBP_CELL_CYCLE
F
ICOSLG target genes NicheNet regulatory potential
p-value=3.4x10-20
0.0020 0.0015 0.0010 0.0005 0.0000
Th_m5 genes
Genes from all other modules
G
ICOSLG interaction upregulated in Factor 2-high donors
ICOSLG
ICOS
cMono
Th
MHC Class II
TCR
Increased cell-cycle, MAPK activation, etc.
Contribution to nephritis
H
Padj = 2.4x10-15
Th_m9 module expression
THBS1 expression in cDC
0.3 0.2 0.1 0.0 −0.1
−1 0 1 2 3
Padj = 1.6x10-14
THBS1 expression in cDC
Factor 3 donor score
3 2 1 0 −1
−0.1 0.0 0.1 0.2

Figure 4. LR approach overview and inference of interactions in SLE multicellular patterns.

(A) LR interaction inference model. The bottom text describes the linear model used to test whether ligand expression in a source cell type is significantly associated with a co-expression gene module in a receptor-bearing cell type across donors (Methods).

(B) Results of the scITD LR analysis using the CellChat LR pair database (left). Rows are ligand hits from various source cell types. Columns are gene co-expression modules from various target cell types. Rows are grouped by source cell type and columns are grouped by target cell type. Values in the main body of the heatmap indicate adjusted p-values for ligand-module associations. Only the top significant results are shown (Methods). Rows and columns are clustered within each block. Also shown are ligand-factor association adjusted p-values (right). Arrows highlight a single ligand-module combination that is displayed in more detail in (C).

(C) Association between gene module B_m1 and expression of ligand *TNFSF13B* in cMonocytes (left). Association between *TNFSF13B* expression in cMonocytes with donor scores for factor 1 (right). The line is a linear model (also applies to D and H).

(D) Association between gene module Th_m5 and expression of ligand *ICOSLG* in cMonocytes (left). Association between *ICOSLG* expression in cMonocytes with donor scores for factor 2 (right).

(E) Enriched gene sets in module Th_m5. Adjusted p-values are shown in green boxes. Enrichment was tested using the hypergeometric test with gene sets from GOBP and BioCarta. Only results with adjusted p-values less than 0.005 are shown.

(F) ICOSLG-target gene regulatory potential scores (from NicheNet) are shown for genes in the Th_m5 module or all other modules. The Wilcoxon rank-sum test was used to test for a difference in medians between the two groups.

(G) Diagram of ICOSLG-ICOS interaction.

(H) Association between gene module Th_m9 and expression of ligand *THBS1* in cDCs (left). Association between *THBS1* expression in cDCs with donor scores for factor 3 (right).

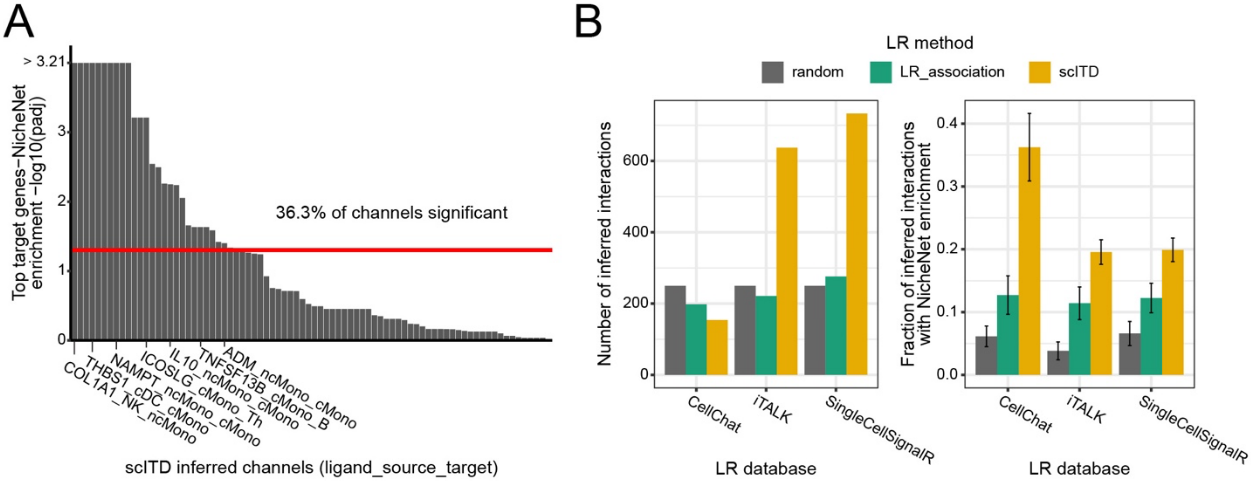


Figure S5. Validation of the scITD LR inference approach using NicheNet regulatory potential scores.

(A) LR channels inferred using scITD with the CellChat LR pair database. The height of each bar indicates the adjusted p-value (calculated with the Wilcoxon rank-sum test) comparing NicheNet regulatory potential for the top 200 potential target genes to the rest (Methods). The red bar indicates an adjusted p-value of 0.05.

(B) The number of inferred LR channels using scITD compared to a simple LR pair association approach (left). The plot on the right shows the fraction of inferred LR channels with significant NicheNet enrichment (as calculated in G). The "random" method is a random selection of 250 LR channels as a background comparison. Error bars are the standard error of the mean.

Factors often involve shifting cell subtype compositions across donors

Different scenarios may underlie the inter-individual transcriptional variation captured by the scITD factors. The variation may represent differences in gene expression within a given cell type (e.g., different activation states of CD4 T cells), or it may reflect altered proportions of cell subtypes (e.g., regulatory versus effector CD4 T cells) in different samples. We implemented a strategy to identify cell subtype "compositional shifts" in an automated fashion (Figure 5A). For a given major cell type, we first perform subclustering to various resolutions. As finer resolution clusters may capture disease-specific expression states, we generally check that non-canonical subtypes are present to a minimal extent in the samples from healthy donors. For a given subclustering, we can then test whether the estimated cell subtype proportions are significantly associated with a given factor. This is demonstrated with the major Th cell cluster using the SLE-only portion of the dataset (Figure 5B). We found that certain subclustering resolutions yielded significant associations with the factor scores, indicating that compositional shifts likely contribute to the observed inter-individual variation.

We then selected a single subclustering resolution for each major cell population to investigate cell subtypes further (Figure 5C, 5D, S6A, and S6B). Firstly, we observed that donors with high factor 1 scores (those who had high ISG expression) tended to have lower fractions of the Th_1 subtype (Figure 5E and 5F). This subtype overexpressed genes such as *CCR7* and *SELL* indicating that it likely represents naïve Th cells (Figure 5D left). This result is consistent with the aforementioned shift in naïve Th cells when using the previously ascribed subtype annotations. Factor 1 donor scores were also significantly associated with altered proportions of B cell subtypes (Figure 5G). Specifically, the ISG expressing patients appeared to have expanded populations of activated memory B cells and transitional B cells. Markers for the activated memory B cells, B_4 (Figure 5D right), match those from a recent single-cell study of healthy individuals (King et al., 2021). The transitional B cell subtype (B_3) was annotated by marker genes, *CD72* and *TCL1A*, identified from previous studies (Shen et al., 2020; Stewart et al., 2020). These findings are also consistent with our observation of *TNFSF13B* upregulation in these patients (identified in our LR analysis), as BAFF has been shown to enhance survival of these B cell subtypes (Hsu et al., 2002). Transitional B cells have previously been shown to be at increased proportions in SLE patients and they are known to play a role in regulating Th differentiation and reducing their proliferation (Simon et al., 2016; Dieudonné et al., 2019; Liu et al., 2019; Zhou et al., 2020). Along these lines, we expected to see reduced Th numbers overall. Therefore, we tested whether factor 1 donor scores were also associated with the composition of the major cell types. We observed a significant overall association, with the Th cells at reduced proportions for the SLE patients with high ISG expression (Figure 5H). Analysis of cell subtype compositional shifts with the other factors also yielded several significant associations with relevant connections to their corresponding multicellular patterns (Figure S6C-G).

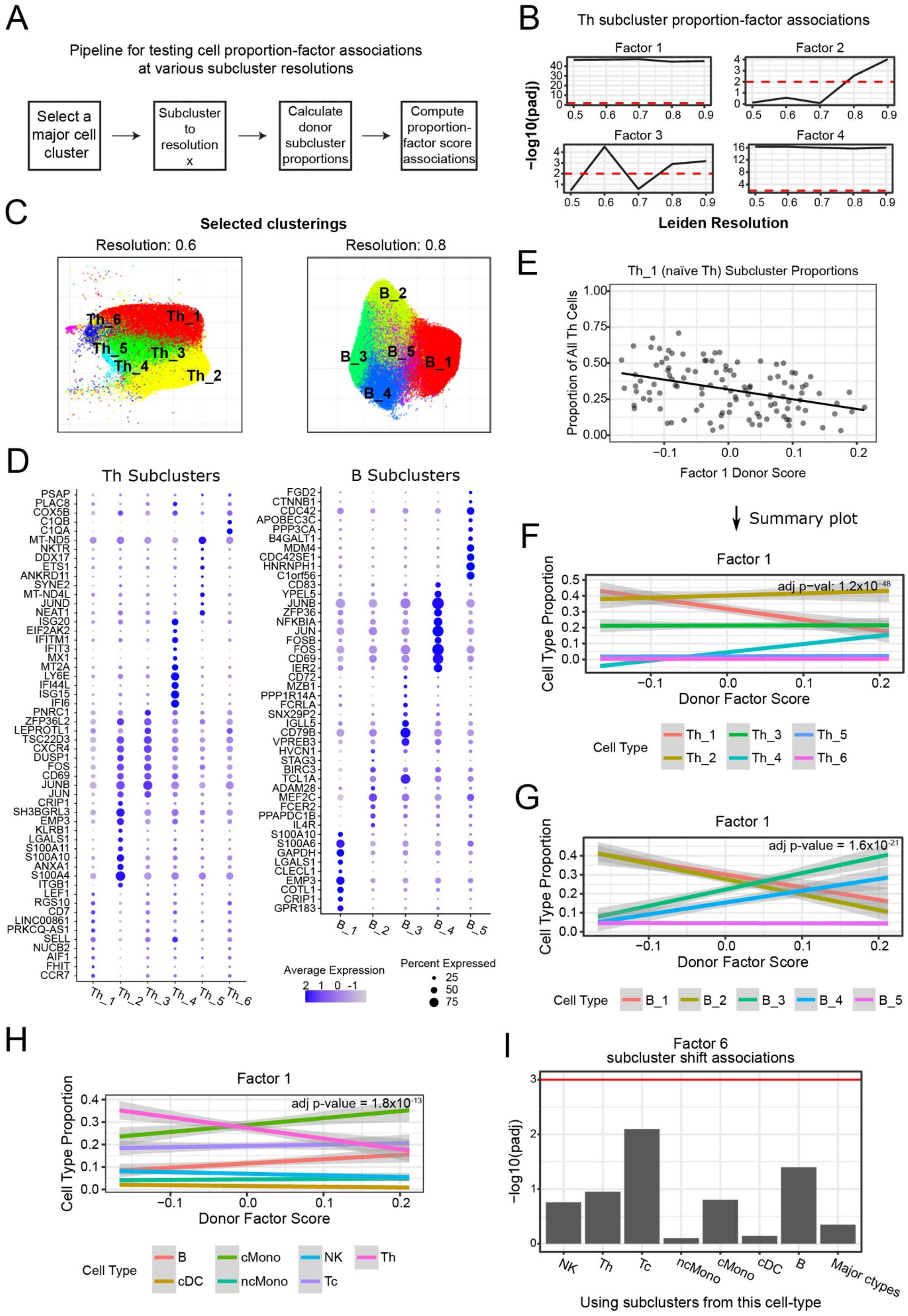

A
Pipeline for testing cell proportion-factor associations at various subcluster resolutions
Select a major cell cluster
Subcluster to resolution x
Calculate donor subcluster proportions
Compute proportion-factor score associations
B
Th subcluster proportion-factor associations
Factor 1
Factor 2
Factor 3
Factor 4
−log10(padj)
Leiden Resolution
0.5 0.6 0.7 0.8 0.9
C
Selected clusterings
Resolution: 0.6
Resolution: 0.8
Th_1
Th_2
Th_3
Th_4
Th_5
Th_6
B_1
B_2
B_3
B_4
B_5
D
Th Subclusters
B Subclusters
PSAP PLAC8 COX5B C1QB C1QA MT-ND5 NKTR DDX17 ETS1 ANKRD11 SYNE2 MT-ND4L JUND NEAT1 ISG20 EIF2AK2 IFITM1 IFIT3 MX1 MT2A LY6E IFI44L ISG15 IFI6 PNRC1 ZFP36L2 LEPROTL1 TSC22D3 CXCR4 DUSP1 FOS CD69 JUNB JUN CRIP1 SH3BGRL3 EMP3 KLRB1 LGALS1 S100A11 S100A10 ANXA1 S100A4 ITGB1 LEF1 RGS10 CD7 LINC00861 PRKCQ-AS1 SELL NUCB2 AIF1 FHIT CCR7
FGD2 CTNNB1 CDC42 APOBEC3C PPP3CA B4GALT1 MDM4 CDC42SE1 HNRNPH1 C1orf56 CD83 YPEL5 JUNB ZFP36 NFKBIA JUN FOSB FOS CD69 IER2 CD72 MZB1 PPP1R14A FCRLA SNX29P2 IGLL5 CD79B VPREB3 HVCN1 STAG3 BIRC3 TCL1A ADAM28 MEF2C FCER2 PPAPDC1B IL4R S100A10 S100A6 GAPDH LGALS1 CLECL1 EMP3 COTL1 CRIP1 GPR183
Th_1 Th_2 Th_3 Th_4 Th_5 Th_6
B_1 B_2 B_3 B_4 B_5
Average Expression
2 1 0 -1
Percent Expressed
25
50
75
E
Th_1 (naïve Th) Subcluster Proportions
Proportion of All Th Cells
Factor 1 Donor Score
Summary plot
F
Factor 1
adj p-val: 1.2x10^-48
Cell Type Proportion
Donor Factor Score
Cell Type
Th_1 Th_3 Th_5
Th_2 Th_4 Th_6
G
Factor 1
adj p-value = 1.6x10^-21
Cell Type Proportion
Donor Factor Score
Cell Type B_1 B_2 B_3 B_4 B_5
H
Factor 1
adj p-value = 1.8x10^-13
Cell Type Proportion
Donor Factor Score
Cell Type
B cMono NK Th
cDC ncMono Tc
I
Factor 6
subcluster shift associations
-log10(padj)
NK Th Tc ncMono cMono cDC B Major ctypes
Using subclusters from this cell-type

Figure 5. Identifying cell-type-composition associations with factors.

(A) Pipeline for testing cell subtype compositional shifts at various subclustering resolutions.

(B) Th subtype composition-factor associations for various subclustering resolutions. The p-values are calculated using multiple linear regression with an F-test. The response variable is the factor donor score and the regressors are balances calculated from the cell subtype proportions for each donor (Methods).

(C) UMAP plots of subclusters for Th cells (left) and B cells (right) that were used in further analyses.

(D) Marker genes for Th cell subclusters (left) and B cell subclusters (right). Marker genes were determined from DE tests between subclusters using Conos (Barkas et al., 2019).

(E) Th_1 subcluster (naïve Th cells) proportions plotted against factor 1 donor scores. The line is a linear model.

(F) Associations between all Th cell subcluster proportions and factor 1 donor scores. The standard error is used for the error bounds, and the p-value is calculated as in (B). This also applies to (G-I).

(G) Association between factor 1 donor scores with proportions of B cell subclusters.

(H) Associations between factor 1 donor scores and donor proportions of the major cell types.

(I) Cell subtype proportion associations with factor 6, the sex-associated factor. The red bar indicates a .001 adjusted p-value.

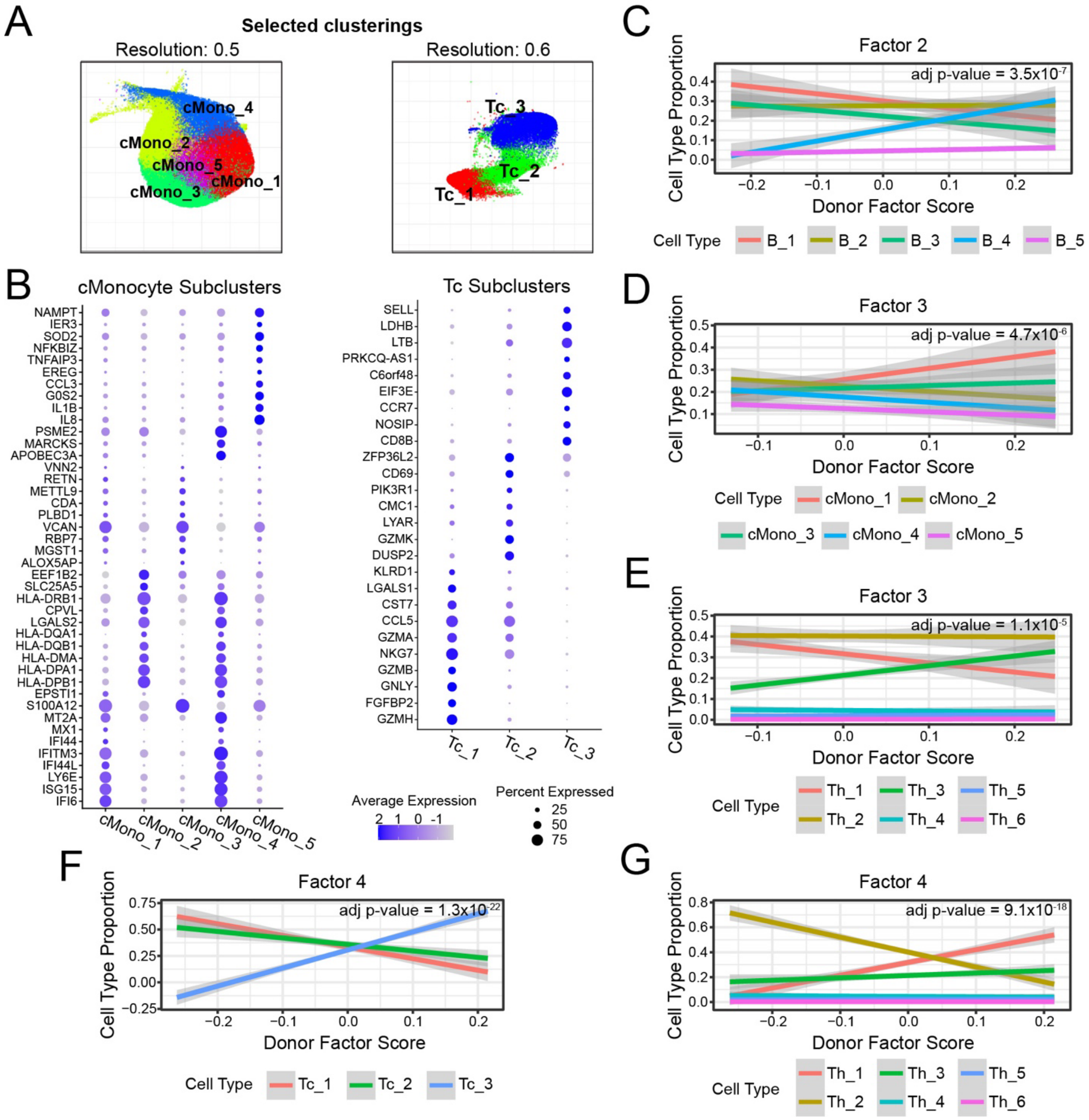


Figure S6. Additional cell composition analysis.

(A) UMAP plots of subclusters for cMonocyte cells (left) and Tc cells (right) that were used in further analyses.

(B) Marker genes for cMonocyte subclusters (left) and Tc cell subclusters (right). Marker genes were determined from DE tests between subclusters using Conos.

(C) Association significance between B cell subcluster proportions and factor 2. The standard error is used for the error bounds, and the p-value is calculated as in (Figure 5B). This also applies to (D-G).

(D) Association significance between cMonocyte subcluster proportions and factor 3.

(E) Association significance between Th subcluster proportions and factor 3.

(F) Association significance between Tc subcluster proportions and factor 4.

(G) Association significance between Th subcluster proportions and factor 4.

scITD extracts multicellular patterns of gene expression associated with COVID-19 severity

Next, we applied scITD to analyze a large scRNA-seq dataset consisting of 83 COVID-19 patients and 20 healthy controls (Stephenson et al., 2021). The patients demonstrated varying degrees of disease severity at the time of sample collection, ranging from asymptomatic to critical. We again limited our analysis to the major cell populations (Figure 6A). Similar to our SLE analysis, the decomposition of this data yielded one factor characterized by pan-cell type ISG expression (factor 1) (Figure 6C). High ISG expression was seen in a subset of the COVID-19 patients but not in the healthy controls (Figure 6D). Interestingly, the patients with critical disease also had significantly lower ISG expression compared to other patients (Figure 6D). This may be partially due to the presence of anti-interferon autoantibodies as indicated by recent reports (Bastard et al., 2020; Wang et al., 2021; van der Wijst et al., 2021).

As ISG seems to play opposite roles in COVID-19 (protective role) compared to SLE (pathogenic role), we aimed to identify differences in ISG expression between acute viral infection and autoimmunity. In comparing the ISG multicellular patterns extracted from the COVID-19 dataset versus the SLE dataset, we found many overlapping significant genes per cell type (Figure 6E, S7A, and S7B), and the most strongly associated canonical ISGs (e.g., IFI6, MX1, XAF1, etc.) were all significant in both datasets (Figure 6E and S7A). However, the ISG multicellular pattern in the SLE dataset contained many genes that were not significant in the COVID dataset (Figure S7B). These SLE-specific factor 1 genes were enriched for biological processes such as ATP metabolic process (Tc cells), cell activation (NK cells), and cell-cell adhesion (NK cells) among others (Figure 6F). Enrichment of ATP metabolic processes in Tc cells is particularly intriguing as one recent SLE study demonstrated striking metabolic changes in Tc cells that take place only in response to chronic interferon exposure and were absent with short-term exposure (Buang et al., 2021). As a negative control, we compared the ISG factor of this COVID-19 dataset to that of another COVID-19 scRNA-seq dataset (van der Wijst et al., 2021) and observed no significant gene sets among the dataset-specific genes. We also noticed that many of the SLE-unique factor 1 genes represented components of the proteasome, including the immunoproteasome subunit, *PSMB10,* upregulated in all cell types analyzed. While proteasome subunits are significantly upregulated with ISGs in both COVID-19 and SLE (Figure 2D and 6C), there appears to be a far stronger enrichment of these genes in the SLE ISG factor (Figure 6G). This was observed for Tc cells, Th cells, and NK cells (Figure S7C). Two genetic studies of ISG expression in SLE found a significant variant in *PPM1H*, a gene that can remove proteolysis signals (Kariuki et al., 2010; Ghodke-Puranik et al., 2020). Therefore, genetics or other disease-specific factors may play a role in differentiating the ISG multicellular pattern in SLE from that seen in other illnesses.

Unlike the ISG factor, factor 2 of the COVID-19 decomposition significantly stratified patients along a continuous spectrum disease severity (Figure 7A). The multicellular pattern for this factor consisted of multiple cell-type-specific biological processes including cell-cycle (NK and T cells) and various signaling cascades (in cMonocytes, NK, and T cells) among others (Figure 7B). We further examined this factor for associations with cell proportions. We found that the more severe patients had reduced proportions of activated Th cells (IL22+) as well as increased proportions of terminal effector Tc cells (Figure S7D). Similar subcluster proportion associations were also reported in the original study that generated this dataset (Stephenson et al., 2021). We also found that the factor 2-high donors had significantly increased proportions of proliferating Th cells, Tc cells, and NK cells (Figure S7D). This is consistent with the observed enrichment of cell-cycle processes in these cell types for these patients (Figure 7B).

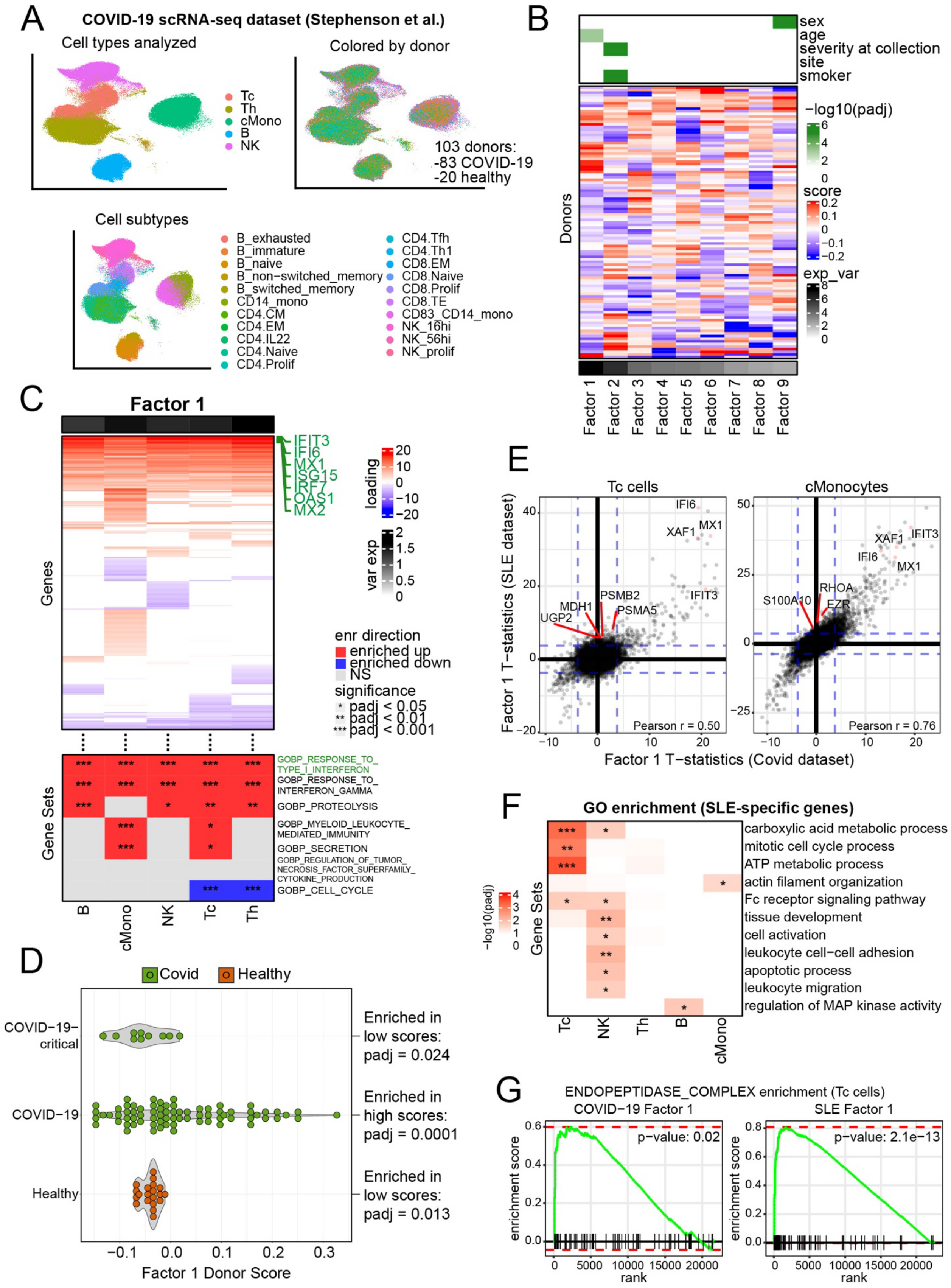

A
COVID-19 scRNA-seq dataset (Stephenson et al.)
Cell types analyzed
Tc
Th
cMono
B
NK
Colored by donor
103 donors:
-83 COVID-19
-20 healthy
Cell subtypes
B_exhausted
B_immature
B_naive
B_non-switched_memory
B_switched_memory
CD14_mono
CD4.CM
CD4.EM
CD4.IL22
CD4.Naive
CD4.Prolif
CD4.Tfh
CD4.Th1
CD8.EM
CD8.Naive
CD8.Prolif
CD8.TE
CD83_CD14_mono
NK_16hi
NK_56hi
NK_prolif
B
sex
age
severity at collection
site
smoker
Donors
−log10(padj)
6
4
2
0
score
0.2
0.1
0
−0.1
−0.2
exp_var
8
6
4
2
0
Factor 1
Factor 2
Factor 3
Factor 4
Factor 5
Factor 6
Factor 7
Factor 8
Factor 9
C
Factor 1
IFIT3
IFI6
MX1
ISG15
IRF7
OAS1
MX2
loading
20
10
0
−10
−20
var exp
2
1.5
1
0.5
0
Genes
enr direction
enriched up
enriched down
NS
significance
* padj < 0.05
** padj < 0.01
*** padj < 0.001
Gene Sets
GOBP_RESPONSE_TO_TYPE_I_INTERFERON
GOBP_RESPONSE_TO_INTERFERON_GAMMA
GOBP_PROTEOLYSIS
GOBP_MYELOID_LEUKOCYTE_MEDIATED_IMMUNITY
GOBP_SECRETION
GOBP_REGULATION_OF_TUMOR_NECROSIS_FACTOR_SUPERFAMILY_CYTOKINE_PRODUCTION
GOBP_CELL_CYCLE
B
cMono
NK
Tc
Th
D
Covid
Healthy
COVID−19−critical
COVID−19
Healthy
Enriched in low scores: padj = 0.024
Enriched in high scores: padj = 0.0001
Enriched in low scores: padj = 0.013
−0.1
0.0
0.1
0.2
0.3
Factor 1 Donor Score
E
Tc cells
cMonocytes
Factor 1 T−statistics (SLE dataset)
Factor 1 T−statistics (Covid dataset)
IFI6
MX1
XAF1
IFIT3
PSMB2
MDH1
UGP2
PSMA5
Pearson r = 0.50
IFIT3
XAF1
IFI6
MX1
RHOA
S100A10
EZR
Pearson r = 0.76
40
20
0
−20
50
25
0
−25
−10
0
10
20
F
GO enrichment (SLE-specific genes)
carboxylic acid metabolic process
mitotic cell cycle process
ATP metabolic process
actin filament organization
Fc receptor signaling pathway
tissue development
cell activation
leukocyte cell−cell adhesion
apoptotic process
leukocyte migration
regulation of MAP kinase activity
−log10(padj)
4
3
2
1
0
Gene Sets
Tc
NK
Th
B
cMono
G
ENDOPEPTIDASE_COMPLEX enrichment (Tc cells)
COVID−19 Factor 1
SLE Factor 1
p−value: 0.02
p−value: 2.1e−13
enrichment score
rank
0
5000
10000
15000
20000

Figure 6. scITD extracts an ISG pattern from a COVID-19 dataset.

(A) UMAP plot of single-cell gene expression from Stephenson et al. (2021) colored by the major cell types used in the scITD analysis (top left), cell subtype annotations (bottom), and source donors (top right).

(B) Donor scores plot with metadata associations (top) and ordered by the explained variance of each factor (bottom). Association p-values were calculated using univariate linear model F-tests. Rows are hierarchically clustered.

(C) Factor 1 loadings for significant genes (top) and select enriched gene sets (bottom). Enrichment significance was calculated using the FGSEA R package with GOBP gene sets (also applies to G). Rows of the loadings heatmap are hierarchically clustered. Green-colored genes are the top leading-edge genes identified in the enrichment of the Response to Type I Interferon pathway.

(D) Association between COVID-19 status and factor 1 scores (top). FGSEA running enrichment tests were used to calculate enrichment of patient groups at either end of the factor 1 donor scores.

(E) Comparing gene expression-factor 1 association T-statistics (by linear model) between the COVID-19 dataset and the SLE dataset. Dashed blue lines represent adjusted p-values of 0.01. Gene labels highlight some genes that are highly significant in both datasets as well as some genes that are significant only in the SLE dataset.

(F) GO gene set overrepresentation analysis of SLE-specific ISGs compared to all upregulated ISGs identified from both datasets.

(G) Enrichment of proteosome-component genes (endopeptidase complex gene set) in ISG-high patients of either the COVID-19 decomposition (left) or the SLE decomposition (right). FGSEA was used to calculate p-values.

By further applying our LR inference technique, we identified a strong candidate LR interaction connected to this multicellular pattern. The interaction included the ligand IL16 expressed from Th cells interacting with the CD4 receptor on cMonocytes. Specifically, we observed high correlations between *IL16* expression in Th cells and various co-expression gene modules in cMonocytes, including cMono_m14 (Figure 7D right). Module cMono_m14 was significantly enriched for MHC Class II genes (Figure 7E), indicating a possible role of Th derived *IL16* in regulating the expression of these genes in monocytes. Corroborating this, earlier studies have shown that IL16 can upregulate monocytic MHC Class II genes, including *HLA-DR* (Cruikshank et al., 1987). Here, we further found Th *IL16* to be downregulated in donors with high factor 2 scores (Figure 7D left), and the factor 2 multicellular pattern shows that these donors also have downregulation of *HLA-DR* expression in monocytes (Figure 7B top). Interestingly, reduced expression of *HLA-DR* in monocytes has also been found in patients with sepsis, a process that shares many of the same pathophysiological features as severe COVID-19 (Winkler et al., 2017; Olwal et al., 2021). Two recent studies have also identified reduced levels of HLA-DR protein on monocytes of critically ill COVID-19 patients (Giamarellos-Bourboulis et al., 2020; Spinetti et al., 2020). Given our result and the prior literature, further studies should be conducted to determine whether reduced *IL16* contributes to increased COVID-19 severity via reduced *HLA-DR* expression.

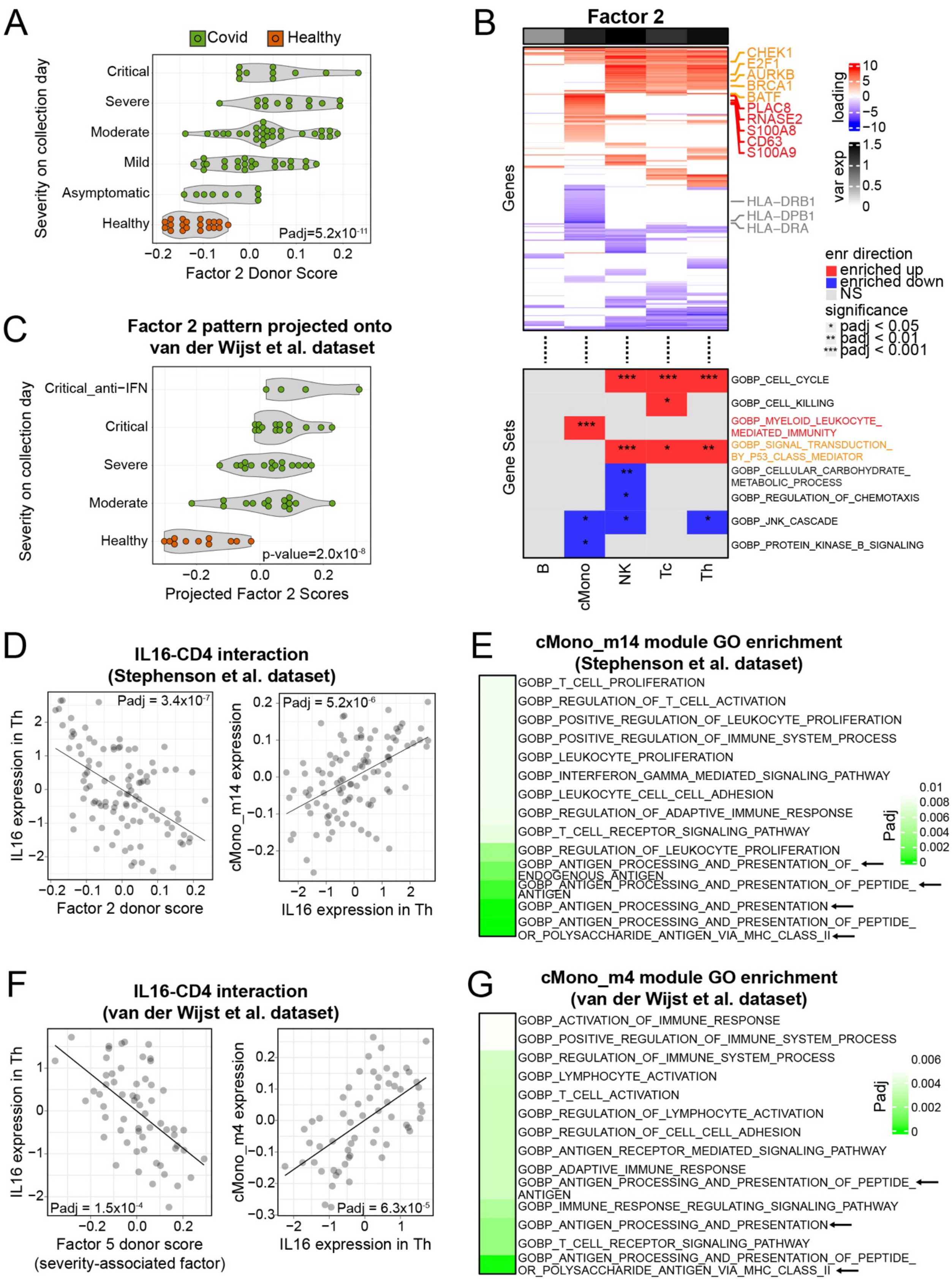

A
Covid
Healthy
Severity on collection day
Critical
Severe
Moderate
Mild
Asymptomatic
Healthy
Padj=5.2x10^-11
−0.2 −0.1 0.0 0.1 0.2
Factor 2 Donor Score
B
Factor 2
Genes
CHEK1
E2F1
AURKB
BRCA1
BATF
PLAC8
RNASE2
S100A8
CD63
S100A9
HLA−DRB1
HLA−DPB1
HLA−DRA
loading
10
5
0
−5
−10
var exp
1.5
1
0.5
0
enr direction
enriched up
enriched down
NS
significance
* padj < 0.05
** padj < 0.01
*** padj < 0.001
Gene Sets
GOBP_CELL_CYCLE
GOBP_CELL_KILLING
GOBP_MYELOID_LEUKOCYTE_MEDIATED_IMMUNITY
GOBP_SIGNAL_TRANSDUCTION_BY_P53_CLASS_MEDIATOR
GOBP_CELLULAR_CARBOHYDRATE_METABOLIC_PROCESS
GOBP_REGULATION_OF_CHEMOTAXIS
GOBP_JNK_CASCADE
GOBP_PROTEIN_KINASE_B_SIGNALING
B
cMono
NK
Tc
Th
C
Factor 2 pattern projected onto van der Wijst et al. dataset
Severity on collection day
Critical_anti−IFN
Critical
Severe
Moderate
Healthy
p-value=2.0x10^-8
−0.2 0.0 0.2
Projected Factor 2 Scores
D
IL16-CD4 interaction (Stephenson et al. dataset)
Padj = 3.4x10^-7
IL16 expression in Th
Factor 2 donor score
Padj = 5.2x10^-6
cMono_m14 expression
IL16 expression in Th
E
cMono_m14 module GO enrichment (Stephenson et al. dataset)
GOBP_T_CELL_PROLIFERATION
GOBP_REGULATION_OF_T_CELL_ACTIVATION
GOBP_POSITIVE_REGULATION_OF_LEUKOCYTE_PROLIFERATION
GOBP_POSITIVE_REGULATION_OF_IMMUNE_SYSTEM_PROCESS
GOBP_LEUKOCYTE_PROLIFERATION
GOBP_INTERFERON_GAMMA_MEDIATED_SIGNALING_PATHWAY
GOBP_LEUKOCYTE_CELL_CELL_ADHESION
GOBP_REGULATION_OF_ADAPTIVE_IMMUNE_RESPONSE
GOBP_T_CELL_RECEPTOR_SIGNALING_PATHWAY
GOBP_REGULATION_OF_LEUKOCYTE_PROLIFERATION
GOBP_ANTIGEN_PROCESSING_AND_PRESENTATION_OF_ENDOGENOUS_ANTIGEN
GOBP_ANTIGEN_PROCESSING_AND_PRESENTATION_OF_PEPTIDE_ANTIGEN
GOBP_ANTIGEN_PROCESSING_AND_PRESENTATION
GOBP_ANTIGEN_PROCESSING_AND_PRESENTATION_OF_PEPTIDE_OR_POLYSACCHARIDE_ANTIGEN_VIA_MHC_CLASS_II
Padj
0.01
0.008
0.006
0.004
0.002
0
F
IL16-CD4 interaction (van der Wijst et al. dataset)
IL16 expression in Th
Padj = 1.5x10^-4
Factor 5 donor score (severity-associated factor)
cMono_m4 expression
Padj = 6.3x10^-5
IL16 expression in Th
G
cMono_m4 module GO enrichment (van der Wijst et al. dataset)
GOBP_ACTIVATION_OF_IMMUNE_RESPONSE
GOBP_POSITIVE_REGULATION_OF_IMMUNE_SYSTEM_PROCESS
GOBP_REGULATION_OF_IMMUNE_SYSTEM_PROCESS
GOBP_LYMPHOCYTE_ACTIVATION
GOBP_T_CELL_ACTIVATION
GOBP_REGULATION_OF_LYMPHOCYTE_ACTIVATION
GOBP_REGULATION_OF_CELL_CELL_ADHESION
GOBP_ANTIGEN_RECEPTOR_MEDIATED_SIGNALING_PATHWAY
GOBP_ADAPTIVE_IMMUNE_RESPONSE
GOBP_ANTIGEN_PROCESSING_AND_PRESENTATION_OF_PEPTIDE_ANTIGEN
GOBP_IMMUNE_RESPONSE_REGULATING_SIGNALING_PATHWAY
GOBP_ANTIGEN_PROCESSING_AND_PRESENTATION
GOBP_T_CELL_RECEPTOR_SIGNALING_PATHWAY
GOBP_ANTIGEN_PROCESSING_AND_PRESENTATION_OF_PEPTIDE_OR_POLYSACCHARIDE_ANTIGEN_VIA_MHC_CLASS_II
Padj
0.006
0.004
0.002
0

Figure 7. Analysis of a COVID-19 severity associated multicellular pattern.

(A) Association between COVID-19 severity at sample collection and factor 2 donor scores from the decomposition of the Stephenson et al. COVID-19 dataset. The significance of the association was calculated with a linear model F-test.

(B) Factor 2 loadings from the decomposition of the Stephenson et al. dataset. The heatmap is limited to significant genes only (top). Also shown are select enriched gene sets (bottom). Rows the loadings heatmap are hierarchically clustered.

(C) Donor scores from projecting the factor 2 pattern onto the van der Wijst et al. dataset. The association p-value was calculated the same way as in (A). The color legend is also the same as in (A).

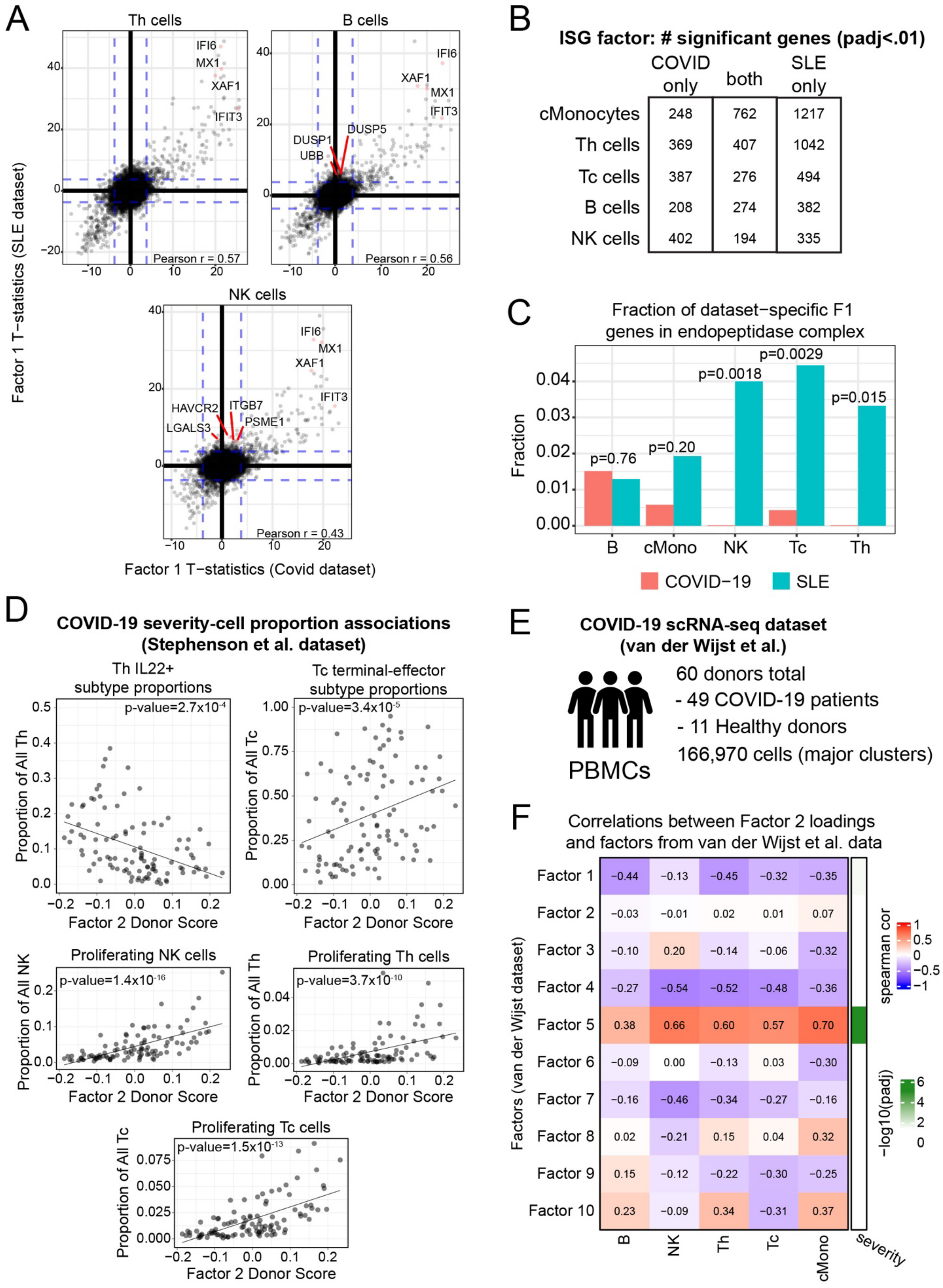
(D) A potential LR interaction between IL16-CD4 (Th cells to monocytes) identified in the Stephenson et al. dataset. Shown is the association between *IL16* expression in Th cells with donor scores for factor 2 (left). Also shown is the association between gene module cMono_m14 and expression of ligand *IL16* in Th cells (right). The line is a linear model.

(E) Top enriched GO gene sets in co-expression module cMono_m14 from the Stephenson et al. dataset (adjusted p-values < 0.009).

(F) The IL16-CD4 interaction identified in the van der Wijst et al. dataset. Shown is the association between *IL16* expression in Th cells with donor scores for factor 5 (the COVID-19 severity-associated factor for this dataset) (left). Also shown is the association between gene module cMono_m4 and expression of ligand *IL16* in Th cells (right). The line is a linear model.

(G) Top enriched GO gene sets in co-expression module cMono_m4 from the van der Wist et al. dataset (adjusted p-values < 0.005).

Finally, we aimed to replicate this severity-associated multicellular pattern by analyzing another COVID-19 scRNA-seq dataset. The test dataset (van der Wijst et al., 2021) consisted of a smaller number of donors but included the same cell types and a similar range of disease severity (not including asymptomatic or mild cases) (Figure S7E). We projected the factor 2 pattern extracted from the Stephenson et al. dataset onto the van der Wijst et al. dataset and similarly found that this pattern significantly stratified patients by disease severity (Figure 7C). A meta-analysis combining the result from these two datasets yielded a Fisher p-value of $4.3x10^{-17}$. The severe and critical patients in the test dataset also had significantly reduced *IL16* expression in their Th cells (Figure 7F left) matching our observation from the Stephenson et al. data. Similarly, *IL16* was again positively associated with a co-expression module in cMonocytes (Figure 7F right) that was enriched for MHC Class II genes (Figure 7G). The consistency of these results across independently generated datasets bolsters our confidence in the connection between this multicellular pattern and COVID-19 disease severity and highlights the conserved nature of a potentially crucial cell-cell interaction.



ISG factor: # significant genes (padj<.01)

	COVID only	both	SLE only
cMonocytes	248	762	1217
Th cells	369	407	1042
Tc cells	387	276	494
B cells	208	274	382
NK cells	402	194	335

	B	NK	Th	Tc	cMono
Factor 1	−0.44	−0.13	−0.45	−0.32	−0.35
Factor 2	−0.03	−0.01	0.02	0.01	0.07
Factor 3	−0.10	0.20	−0.14	−0.06	−0.32
Factor 4	−0.27	−0.54	−0.52	−0.48	−0.36
Factor 5	0.38	0.66	0.60	0.57	0.70
Factor 6	−0.09	0.00	−0.13	0.03	−0.30
Factor 7	−0.16	−0.46	−0.34	−0.27	−0.16
Factor 8	0.02	−0.21	0.15	0.04	0.32
Factor 9	0.15	−0.12	−0.22	−0.30	−0.25
Factor 10	0.23	−0.09	0.34	−0.31	0.37

Figure S7. Additional details for COVID-19 analysis and ISG factor comparison.

(A) Comparing gene expression-factor 1 association T-statistics (by linear model) between the COVID-19 dataset and the SLE dataset. Dashed blue lines represent adjusted p-values of 0.01. Gene labels highlight some genes that are highly significant in both datasets as well as some genes that are significant only in the SLE dataset.

(B) The number of shared and dataset-specific factor 1 associated genes (includes positive and negative associations) between the COVID-19 dataset and the SLE dataset.

(C) Enrichment of endopeptidase complex genes among SLE-specific factor 1 significant genes compared to COVID-19-specific factor 1 genes. The hypergeometric test was used to calculate p-values

(D) Factor 2 associations with cell subtype proportions for IL22+ Th cells (top left) and terminal effector Tc cells (top right). Also showing subtype proportion associations for proliferating cell populations with factor 2 (bottom). The p-values were calculated using linear model F-tests after transforming the proportions to balances (Methods).

(E) Structure of a COVID-19 scRNA-seq dataset by van der Wijst et al. that we used for validating the factor 2 multicellular pattern identified in the Stephenson et al. dataset.

(F) Correlations between loadings from factor 2 of the Stephenson et al. dataset to loadings from factors of van der Wijst et al. dataset decomposition. The right annotation shows the factor-severity association significance for the van der Wijst et al. decomposition. The significance p-values of the disease severity associations were calculated using linear model F-tests.

Discussion

Transcriptional response of complex biological tissues to different conditions or diseases can involve coordinated changes across multiple cell types. These multicellular events may represent the joint reactions of multiple cell types to a common set of stimuli as well as the cell-cell interactions that contribute to the associated phenotypes. Here, we developed a single-cell computational tool, scITD, to identify coordinated multicellular patterns of expression variation across individuals. Our approach can be applied to any scRNA-seq dataset with multiple samples, though it is best geared for use with datasets consisting of many source donors. We validated that the tool can extract meaningful and accurate multicellular patterns using simulated data as well as data from an *in vitro* IFN-beta stimulation experiment. We also developed a component of our method to infer cell-cell interactions and showed that it outperformed a widely used LR inference strategy.

The primary use cases for scITD involve the study of inter-individual variation within different types of sample collections, including those sampling normal individuals, case-control studies, or disease subtyping studies. In this study, we principally applied the tool to discover multicellular patterns that stratify SLE patients. From this dataset, we extracted a multicellular pattern involving ISG expression that separated SLE patients from the healthy controls. This pattern was significantly associated with the presence of autoantibodies and SLE disease activity, recapitulating similar associations from previous studies. Our method uniquely allowed us to identify additional cell-type-specific biological processes co-occurring with ISG expression (including B cell activation, Treg expansion, etc.) suggestive of interactions between circulating cell types. By further connecting the other factors to clinical metadata, we were able to identify a multicellular pattern that was linked to increased frequency of lupus nephritis among the patients positive for anti-dsDNA autoantibodies. While several of the cell-type-specific components of this multicellular pattern have been previously validated and linked to lupus nephritis (e.g., P38 MAPK signaling and ICOSLG signaling from myeloid cells to T cells), our analysis suggests that these pathways may be important in explaining why anti-dsDNA autoantibodies are necessary but insufficient to cause these severe symptoms in SLE.

We further showed that our method can yield insightful patterns in other large patient cohorts by applying it to a COVID-19 scRNA-seq dataset. We extracted a factor consisting of pan-cell-type ISG expression that separated healthy donors from a subset of COVID-19 patients. This ISG multicellular pattern appeared to play a protective role in the disease as the critically ill patients had reduced ISG expression. We compared this ISG pattern to that from the SLE dataset and identified cell-type-specific differences that may be connected to the chronic nature of SLE. Another factor from the COVID-19 dataset was strongly associated with the continuous spectrum of disease severity and displayed a more complex multicellular pattern of expression. We demonstrated that this pattern was also predictive of COVID-19 severity in an independently generated scRNA-seq dataset, highlighting the robustness and generalizability of the extracted multicellular patterns. Lastly, we highlighted a high-confidence LR interaction involving the ligand IL16 expressed in T cells and interacting with the CD4 receptor on cMonocytes. This interaction appeared to be downregulated for the most severe COVID-19 patients, highlighting a need for follow-up studies elucidating its role in this disease and evaluating its therapeutic potential.

Finally, this work can be extended in several directions. For one, scITD can be applied to more thoroughly investigate how different technologies, processing techniques, and disease models impact expression jointly within each cell type. We briefly demonstrated how scITD can be used in this way, showing that ambient RNA within 10X Chromium lanes tends to alter expression across all cell types. More work in this area could lead to better-informed designs of batch-correction methods tailored for various scenarios. In another application area, it may also be valuable to connect the scITD multicellular patterns with genetic variants. By further applying techniques such as Mendelian randomization, it will be possible to test causal hypotheses for genes that may be upstream drivers of the observed patterns. Another use case for the scITD is in the study of multi-tissue patterns of gene expression (as opposed to multi-cell type patterns). Since scITD uses scRNA-seq data at the pseudobulk-level, the tool can be directly applied to bulk RNA-seq datasets generated from multiple donors with multiple tissue types (e.g., the GTEx studies). This could allow one to connect gene expression changes in the blood with the expression states of less accessible tissues. Overall, scITD offers a novel approach to single-cell data analysis, extending our ability to study the complex biological processes that stratify individuals in health and disease.

Methods

scITD pipeline and details

The first step is to transform a gene-by-cells UMI counts matrix into a pseudobulked tensor for decomposition. For cells of a given sample and cell type, all counts for each gene are summed and divided by the total counts from all genes. Then, the trimmed-mean of M values (TMM) method in edgeR (Robinson et al., 2010) is used to adjust library sizes of the pseudobulked counts and the data are normalized and log-transformed. Before this previous step, we only retain donors with at least a minimum number of cells in all cell types, as the Tucker decomposition does not allow for NA values in the tensor. Next, we then compute a measure of normalized variance for each gene, as is calculated in the R package, pagoda2 (Barkas et al., 2021). Overdispersed genes are selected from each cell type and each gene-by-sample matrix is reduced to the union of overdispersed genes from all cell types. After normalization, the data are centered and unit scaled across samples. We then rescale genes by their normalized variance calculated previously. This is done by multiplying the expression by the normalized variance value to some power. The power should be set to 0.5 for the resulting variance to equal the normalized variance. We note that increasing the value of the power slightly (often between 1-2) can sometimes improve the quality of the decomposition. Finally, the pseudobulked cell-type matrices are stacked together to form the tensor.

Next, we apply the Tucker tensor decomposition to the resulting tensor. For this, we use the R package, rTensor (Li et al., 2018), which implements Higher-Order Orthogonal Iteration (HOOI) to compute the Tucker decomposition. The formal algorithmic procedure for HOOI can be viewed in the following publication (Sheehan and Saad, 2007). For our case with three dimensions, HOOI outputs three separate factor matrices and a core tensor that can be multiplied together to reconstruct the approximation of the starting tensor. The three separate factor matrices are of dimensions donors-by-donor factors, genes-by-gene factors, and cell-types-by-cell-type factors. The core tensor is of dimensions donor factors-by-gene factors-by-cell-type factors. The standard data reconstruction using these objects is as follows:

$$X \approx G \times_1 A \times_2 B \times_3 C \tag{1}$$

Here, X is the reconstructed tensor, A is the donor factor matrix, B is the gene factor matrix, C is the cell type factor matrix, and G is the core tensor. The operator $\times_n$ indicates multiplication of the matrix on the right side of the operator by the tensor on the left side of the operator along the n^th^ mode of the tensor. We then rearrange the terms in the above equation to yield a “donor-centric” view of the decomposition. To do this, we compute a loadings tensor by multiplying the core tensor only by the gene factor- and cell type factor matrices. This is valid to do because the order of multiplication does not matter when reconstructing the data, as long as the multiplication mode also changes accordingly (Kolda and Bader, 2009). This reordering and simplification appear as follows:

$$\begin{aligned} X &\approx G \times_1 A \times_2 B \times_3 C \\ &= G \times_2 B \times_3 C \times_1 A \\ &= (G \times_2 B \times_3 C) \times_1 A \\ &= F \times_1 A \end{aligned} \tag{2}$$

Here, F is the loadings tensor of dimensions donor factors-by-genes-by-cell types, the rest of the terms are the same as in the previous equation. In Figure 1B, this reconstruction is shown backward as $A \times_1 F$ simply to demonstrate how the tensor times matrix multiplication yields the reconstructed tensor of correct dimensions. As noted previously, the Tucker decomposition to a given number of factors does not have one unique solution. The factor matrices can be rotated by any non-singular square matrix, and as long as the core tensor is counter-rotated by the inverse of the rotation matrix, the reconstruction error will remain unchanged. The core tensor can also be rotated similarly as long as the factor matrices are counter-rotated accordingly. Taking advantage of this property, several groups have found that rotating the factor matrices with independent component analysis (ICA) can improve the interpretability of the factors (Bro; Unkel et al., 2011; Zhou and Cichocki, 2012). Here, we explored the use of various rotations. One such approach we tested was applying ICA rotation to the donor scores matrix, and counter-rotating the core tensor before generating the final donor-centric view of the decomposition. As a note, we needed to normalize the rotated donor scores matrix because ICA does not preserve lengths. After rotating the donor scores matrix, the core tensor is counter-rotated as follows:

$$\hat{G} = X \times_1 \hat{A}^T \times_2 B^T \times_3 C^T \tag{3}$$

Here, $\hat{G}$ is the new core tensor, $\hat{A}^T$ is the transpose of the ICA rotated (normalized) donor factor matrix, C^T is the transpose of the cell type factor matrix, B^T is the transpose of the gene factor matrix. Then, we substitute $\hat{G}$ for G when calculating the loadings tensor F in equation 2.

The primary approach used in most analyses of this study involves a two-step rotation procedure which is a hybrid of ICA and varimax applied to the distinct components that make up the loadings tensor. The intuition behind this approach is to create a core tensor, where each donor factor represents some combination of biologically distinct gene sets (gene factors) in the different cell types. Furthermore, donor factors of the core tensor should be rotated to a simple structure to make the multicellular patterns more modular. This helps to ensure that each gene factor only partakes primarily in one donor factor. This can be achieved by applying varimax to the core tensor. However, it is also necessary to ensure that all gene factors are independent of one another. Otherwise, the optimized core tensor may still yield donor factor loadings with similar sets of relevant genes. To accomplish this, we apply rotations in two separate steps. In the first step, we apply ICA applied to the gene factor matrix and counter-rotate the core tensor (equation 4A). In the second step, we optimize the core tensor by the varimax rotation and counter-rotate the donor matrix (equation 4B). This is calculated as follows:

Step 1
$$\hat{G} = X \times_1 A^T \times_2 \hat{B}^T \times_3 C^T \tag{4A}$$

Step 2
$$\begin{aligned} X_{(1)} &\approx A * R^{T^{-1}} * R^T * \hat{G}_{(1)} * \left(C \otimes \hat{B}\right)^T \\ &= \left(A * R^{T^{-1}}\right) * \left[R^T * \hat{G}_{(1)} * \left(C \otimes \hat{B}\right)^T\right] \\ &= \hat{A} * \hat{F}_{(1)} \end{aligned} \tag{4B}$$

Here, all variables are as previously described, with the addition of $\hat{B}$, which represents the ICA rotated gene factor matrix, and R which now represents the orthonormal rotation matrix found by varimax when applied to $\hat{G}_{(1)}^T$. The symbol $\otimes$ is the Kronecker product. For this approach, we use the identity matrix for C. For all of our analyses, we also set the ranks of C equal to the number of cell types, since we used a relatively small number of cell types. To project a multicellular pattern onto new data to obtain donor scores (as in Figure 7F), we first compute loadings as $F_{(1)} = \hat{G}_{(1)} * \left(C \otimes \hat{B}\right)^T$ without rotating the core. Then, we calculate the new donor scores by $A_{proj} = X_{(1)} * F_{(1)}^T$. Then, the columns of the new scores are normalized to a magnitude of 1 and are rotated by applying the varimax rotation matrix from the core optimization as $\hat{A}_{proj} = A_{proj} * R^{T^{-1}}$.

For visualization purposes, we order the factors in the donor scores matrix from highest to lowest explained variance. To calculate the variance explained by a given factor, we first compute the reconstructed tensor $\tilde{X}$ using only the selected factor. Then variance explained is simply the calculation for the coefficient of variation:

$$\text{explained variance}_p = 1 - \frac{\|X - \tilde{X}_p\|_F^2}{\|X\|_F^2} \tag{5}$$

The subscript p indicates the factor used for reconstructing the data. The subscript F on the double brackets indicates the Frobenius norm. In the loadings matrices, we also display the amount of variance explained by each cell type component. To compute explained variance for individual cell type components within a factor, all values of the reconstructed tensor for all other cell types not under consideration are set to 0:

$$\text{explained variance}_{p,c} = 1 - \frac{\|X - \tilde{Y}_{p,c}\|_F^2}{\|X\|_F^2}, \tilde{Y}_{i,j,k} = \begin{cases} \tilde{X}_{i,j,k}, & k = c \\ 0, & k \neq c \end{cases} \tag{6}$$

The subscript c refers to the cell type being used for calculating explained variance. The subscripts i, j, k refer to the donor, gene, and cell type index of the tensor, respectively.

Simulation study

To generate the simulations we used the R package, Splatter (Zappia et al., 2017). Specifically, we applied the Splatter to generate four subpopulations for each of the two cell types (Figure 2A). To generate a set of four subpopulations, two separate group simulations were run and concatenated together. Each simulation generated two groups of cells separated by some DE genes. Half of the cells from the first population of the first simulation were matched to those of the first population in the second simulation. Likewise, half of the cells from the second population of the first simulation were matched to the other half of the cells from the first population of the second simulation and so on. We then assigned groups of donors (by those with upregulation of each multicellular pattern) randomly to cells from the specified cell subpopulations.

We ran the data through the standard scITD pipeline. However, in this analysis, we did not reduce the tensor to only the most variable genes, so that it would be possible to use all genes in the AUC calculation. We computed the Tucker decomposition to two donor factors and four gene factors, as suggested by our rank determination method. We then calculated gene significance p-values for each cell type in each factor. This was done using linear model F-tests with expression as the explanatory variable and donor scores as the response variable. This technique is also used in the main dataset analyses below to identify significant genes per factor. Then, we used these p-values to calculate the ROC AUC for predicting ground truth DE genes that distinguish each of the two multicellular patterns. We further subsampled the simulated dataset to varying sizes to determine the sensitivity of our method to a reduced signal-to-noise ratio.

SLE dataset processing

The SLE scRNA-seq dataset was originally demultiplexed using an updated version of *demuxlet (Kang et al., 2018)*, and quality control measures were applied using Scanpy (Wolf et al., 2018) with the default parameters. We further filtered out cells with over 10% of their UMIs attributed to mitochondrial genes. This dataset originally contained over 200 donors with transcriptomes from over 1.2 million cells. To make it possible to use this dataset with scITD in its current framework, we reduced the dataset down to one sample per donor, used only the largest cell clusters, and restricted it to only those donors with at least 20 cells in each major cell cluster. Cell clusters and annotations from the original study were used. This left us with 171 donors and 632,733 cells. The median number of cells per donor for B, NK, Th, Tc, cDC, cMonocytes, and ncMonocytes were 421, 264, 1145, 688, 50, 939, and 144 cells respectively. According to our simulation study discussed previously, these quantities were more than sufficient to extract multicellular patterns with high accuracy. We formed the expression tensor using the standard scITD pipeline, and we also applied ComBat batch correction (Johnson et al., 2007) at the level of 10X lanes to each cell-type slice of the tensor. For our primary SLE analysis, we used the hybrid rotation method and decomposed the data into 7 donor factors and 20 gene factors. We used the same ranks parameters for the decomposition run on only the SLE donors. We used the SLE-only decomposition to compare factor 1 cMonocyte dysregulated genes with DE genes from the IFN-beta experiment. We then computed associations between the donor scores for each factor and the metadata variables displayed in Figure 2C using linear model F-tests with donor scores as the response variable. The statistical tests used for computing associations between factor donor scores and the other clinical variables depended on the type of the variable being tested. For the ordinal variables such as SLEDAI score and SLICC score, we employed the ordinal logistic regression with the "probit" method, and p-values were calculated using the resulting t-statistics. For binary variables such as the presence of symptoms or prednisone use, we employed a logistic regression with a chi-square test for significance. We only tested the binary variables that were present in at least 20 donors. Multiple hypothesis test correction was applied with the BH procedure. To test for the association of factor 2 with the co-occurrence of lupus nephritis and anti-dsDNA autoantibodies, we first removed any donor scores for donors that did not have anti-dsDNA autoantibodies present. Then we used a sliding window of size 19 to calculate the number of donors within the window that also had lupus nephritis. To calculate a p-value we randomly shuffled donor scores and computed the Spearman correlation between the donor scores and the sliding window count. We repeated this procedure 10000 times to generate a null distribution of correlation values. The null distribution was then used to calculate a p-value, by counting the number of null instances with a larger correlation than the one we observed with the unshuffled data.

IFN-beta experiment data processing

We only kept the cell barcodes labeled as singlets and used the cell-type annotations ascribed by the authors who generated the data (Kang et al., 2018). DE analysis results were also used directly from the original paper where the data were generated. We ran the Tucker decomposition to extract two donor factors, four gene factors. We used the ICA rotation method on donor scores for this dataset, although the results were practically identical when obtained using the other rotations as well. The downsampling procedure was performed as described above with the simulated data, except that instead of AUC, we report the Spearman correlation between the loadings and log2FC values from the DE analysis.

Procedure for rank determination

To help determine the appropriate ranks to decompose the tensor to, we developed a method similar to those commonly used with matrix decompositions. The method works by unfolding the starting tensor along a given mode and computing the SVD to an increasing number of factors. The reconstruction error is calculated for each decomposition. Then, we repeat this procedure multiple times but with a shuffled version of the tensor. The shuffling is done by randomly reassigning cells to donors before tensor formation. This procedure allows us to evaluate when the reduction in real explained variance no longer exceeds that of the randomized scenario.

Procedure for stability analysis

To test the stability of our decomposition, we designed a procedure whereby the data tensor is subsampled to some fraction of donors. For our analysis, we subsampled to 85% of the donors. We then recomputed the decomposition to the same number of factors. The new factors found on the subsampled data were then linked back to the original factors by identifying, for each original factor, which new factor had the highest absolute value correlation with it. We repeated this procedure 500 times and report the max correlations for each original factor with the new factors from the subsampled data.

Procedure for GSEA

To compute enriched gene sets among genes prioritized within each cell type for a given factor, we use the R package FGSEA (Korotkevich et al., 2021). In tests of applying GSEA directly to an individual column of a factor loadings matrix, we noticed some spurious results appear as a result of the low-magnitude non-significant loadings being biased toward either positive or negative loadings. To avoid getting these false-positive hits, we compute a new value for each gene. This is calculated as the sum of unit scaled expression values multiplied by donor scores as follows:

$$S_{g,c,p} = \sum_d E_{g,c,d} * F_{p,d} \tag{7}$$

Here, $S_{g,c,f}$ is the new score for gene g in cell type c for factor p. $E_{g,c,d}$ is the scaled expression of gene g in cell type c for donor d. And $F_{p,d}$ is the donor score for factor f and donor d. We then apply GSEA to the new gene scores for each cell-type factor combination and apply BH multiple hypothesis test corrections. This has shown to be much more robust in preventing false-positive enriched gene sets, while still yielding gene sets that are expected based on the significant genes.

Procedure for cell proportion analyses

To systematically identify shifts in subtype composition, we first aligned cells across the 10X lane batch variable using Conos (Barkas et al., 2019). Then we subclustered each major cell population to varying resolutions using the *findSubcommunities* function from Conos. To get marker genes that distinguish the cell subclusters, we used the *getDifferentialGenes* function. Marker genes were plotted using the *DotPlot* function in Seurat (Satija et al., 2015). To test for associations between a factor and cell subtype composition, we calculated the cell proportions for each donor as the number of their cells from a given subcluster divided by the total number of cells from the corresponding major cell cluster. Next, we convert these proportions to balances using the isometric log-ratio transformation (Egozcue et al., 2003). This converts the dependent set of proportions to an independent set of p-1 variables, where p is the number of cell subtypes for a given major cell cluster. By making this conversion, it allows one to use these variables with standard statistical tests that require covariates to be independent. As a note, we also add a pseudocount of 1 to each cell proportion numerator to avoid infinities when calculating balances. Then, we use the balances as explanatory variables in a multiple linear regression against the donor scores of each factor. An F-test is used to determine whether a given cell type composition is significantly associated with a factor. The procedure is the same for computing the significance of factor associations with the overall major cell type composition.

Procedure for LR analysis

For this analysis, we used a database of protein ligands and receptors from CellChat (Jin et al., 2021). We first identified clusters of co-expressed genes in the pseudobulk data for each cell type using the R package WGCNA (Langfelder and Horvath, 2008). Specifically, we used the signed network and TOM similarity matrix with a module tree cut height of 0.25. Then, we calculated the association between each module eigengene and each ligand from the list of cognate LR pairs. This is only done if the receptor was present in the module cell type to some minimal level (expressed in at least 5 cells for each of the top 15% of donors by ligand expression). Filtering was applied such that we only tested ligands where at least 1% of donors have scaled-adjusted expression above 0.2 (the same is applied to both ligands and receptors in the LR association test below). This removed ligands where high expression is observed in only a few donors or if the ligand has low normalized variance across donors. We also only tested for interactions between different cell types, as we would likely find many false-positive associations between ligands and modules found in the same cell type (due to regulation of the ligand by the same upstream transcription factor regulating the module). The association p-values were calculated by a linear model F-tests with the module eigengene as the response variable and the scaled expression of a ligand as the explanatory variable. P-values were adjusted using the BH procedure. When the CellChat database listed multiple receptor components as required for a specific LR interaction, we required all components of the receptor complex to be expressed. The heatmap in Figure 4B only includes rows with at least one adjusted p-value below 5.0x10-11 and columns with at least one adjusted p-value below 0.001 to reduce the number of results for visual purposes. Finally, we computed the overrepresentation of gene sets within gene co-expression modules of interest. We employed the hypergeometric test to determine whether a module contained a significantly higher proportion of genes from specific gene sets compared to the rest of the genes not found in the module but still included in the analysis. For the B_m1 enrichment tests, we included gene sets from GOBP, KEGG, and Reactome databases. For benchmarking our approach, we compared it to a standard LR association method and a random selection of LR channels. The LR association test inferred an LR interaction if the expression of the ligand from one cell type was significantly associated (adjusted p-value < 0.05) with the expression of its cognate receptor in another cell type. For the scITD LR inference method, we used a more stringent adjusted p-value of 0.0001 to make the number of inferred interactions roughly the same for at least one of the LR pair databases (CellChat database in this case). To demonstrate that our method could enrich for a more confident set of LR interactions, we used NicheNet regulatory potential scores for each ligand. For a given LR channel consisting of a ligand, a source cell type, and a target cell type, we first identified the top 200 genes in the target cell type by the highest absolute value Pearson correlation with the ligand's expression in the source cell type. Then, we compared the regulatory potential scores for these top 200 genes with those from all other genes. We used a Wilcoxon rank-sum test to test for differences in regulatory potential scores between these two groups. As a note, we also applied the *normalize_correlation* function from the *spqn* package to correct for gene-gene correlation biases with mean expression (Wang et al., 2020).

Batch factor analyses

To calculate the soup profile for batch dmx_YE_7-19, we calculated the fraction of each gene's UMI counts over the total using all empty droplets with 10 or fewer UMIs. GC content from each gene was retrieved using the EDASeq package in R (Risso et al., 2011). To test whether the upregulated genes of a given batch factor had a significant association with GC content, we used a two-sample t-test comparing GC contents for upregulated genes versus all other genes included in the analysis. For the non-batch factors, we arbitrarily selected the positive loading direction to call upregulated genes. After removing batch effects and retesting for GC associations, we tested both directions for GC content associations.

Factor rotation comparisons

To evaluate the impact of different factor rotations on the output we compared our hybrid rotation loadings to ICA on donor scores. We used the full SLE dataset for this analysis. For the *loadings rotations*, we ran the Tucker decomposition to extract 7 donor factors and 20 gene factors as in our main SLE analysis. For the *donor scores rotation*, we ran the Tucker decomposition to extract 9 donor factors and 20 gene factors because the explained variance was distributed more evenly among the factors for this rotation compared to the *loadings rotation*. To calculate ISG associations with each factor, we calculated the linear model r-squared value for regressing each ISG against the factor donor scores. The core set of ISGs used included *ISG15, IFI27, IRF7, HERC5, LY6E, MX1, OAS2, OAS3, RSAD2, USP18,* and *GBP5*, which were also used to represent ISG expression in a previous study (Davenport et al., 2018). For the ethnicity association strength comparison, we used the factor from each rotation that was most strongly associated with ethnicity. We evaluated the statistical significance of the associations using linear model F-tests.

Stephenson et al. COVID-19 data preprocessing and analysis

We used the quality-controlled data that was made publicly available (Stephenson et al., 2021). A few donors had multiple samples taken at various time points. We kept only the samples labeled with collection day "D0", so that the dataset contained 1 sample per donor. We also excluded patients with other diseases besides COVID-19, and we also excluded the healthy donors given LPS. We used the previous annotations labeled as "full_clustering" and grouped cell subtypes to form the major cell type clusters. Specifically, for B cells we included "B_exhausted", "B_immature", "B_naive", "B_non-switched_memory", and "B_switched_memory". For cMonocytes, we included "CD14_mono" and "CD83_CD14_mono". For Th cells we included "CD4.CM", "CD4.EM", "CD4.IL22", "CD4.Naive", "CD4.prolif", "CD4.Tfh", and "CD4.Th1". For Tc cells, we included "CD8.EM", "CD8.Naive", "CD8.Prolif", and "CD8.TE". For NK cells we included "NK_16hi", "NK_56hi", and "NK_prolif". We also noticed that some of the cells previously labeled as "B cells" clustered with the plasmablasts and expressed the same gene and protein markers as the plasmablasts. Therefore, these were excluded from the B-cell pseudobulks. After removing donors with less than 2 cells in any of the included cell types, we were left with 103 donors and 452,740 cells from the major cell populations used in the analysis. The median number of cells per donor for B, Tc, Th, NK, and cMonocytes were 386, 558, 573, 593, and 717 cells, respectively. The samples were processed at one of three different sites including Cambridge, NCL, and Sanger. Therefore, we applied ComBat batch correction to account for these technical differences as was done with the SLE dataset. We ran the Tucker decomposition to 9 donor factors and 26 gene factors using our hybrid rotation method. To compare the interferon factor from this dataset to that of the SLE dataset, we computed the association T-statistics for all genes that were present in both datasets against the first factors For the genes which were significant only for SLE factor 1 but not COVID factor 1, we computed enriched GO gene sets using the union of significant factor 1 genes from both datasets as the background with a hypergeometric test. We further used the FGSEA package for computing enrichment of donors by status (COVID-19, healthy, or COVID-19 critical) in either high or low factor one scores. For enrichment of critical patients in factor 1, we removed the healthy donors.

van der Wijst et al. COVID-19 data preprocessing and analysis

We used the quality-controlled data that was made publicly available (van der Wijst et al., 2021). We removed cases that were negative for SARS-CoV-2. We only used samples measured at day 0, such that we retained 1 sample per donor. Broad cell clusters and annotations were used from the original study. After removing donors with less than 2 cells in any of the included cell types, we were left with 60 donors and 166,970 cells from the major cell populations used in the analysis. The median number of cells per donor for B, Tc, Th, NK, and cMonocytes were 208, 459.5, 637, 278.5, and 747 cells, respectively. We also applied ComBat batch correction at the level of 10X lanes. Finally, we ran the Tucker decomposition to 10 donor factors and 30 gene factors using our hybrid rotation method. To project the factor 2 pattern from the Stephenson et al. dataset, we used the intersection of genes from the tensors of both datasets and carried out the projection as described above. We used Fisher's method to compute the meta-analysis p-value combining the severity-association test with that of the larger dataset. To compare factor 2 loadings to the factor loadings from the van der Wijst et al. decomposition, we computed Spearman correlations for each cell type separately.

Data Availability

The IFN-beta stimulation dataset can be found in the Gene Expression Omnibus (GEO) at accession number GSE96583. The count-matrices for the SLE scRNA-seq dataset are available in GEO at accession GSE137029. The Stephenson et al. COVID-19 dataset is publicly available at https://www.covid19cellatlas.org/index.patient.html, titled "COVID-19 PBMC Ncl-Cambridge-UCL". The van der Wijst et al. COVID-19 dataset can be found at https://cellxgene.cziscience.com/collections/7d7cabfd-1d1f-40af-96b7-26a0825a306d .

Code Availability

Our computational method, scITD, and its associated tutorials can be found at https://github.com/kharchenkolab/scITD. scITD is also available on The Comprehensive R Archive Network (CRAN) at https://cloud.r-project.org/web/packages/scITD/index.html. The code used to produce all figures in this paper can be found at https://github.com/j-mitchel/scITD-Analysis/tree/main/figure_generation.

Author Contributions

P.V.K. has formulated the study and with J.M. developed the overall approach. J.M. developed the detailed algorithms with advice from P.V.K. with assistance from E.B. J.M., C.J.Y., and P.V.K. worked on interpretation of results, with help from M.G.G., R.K.P., and R.B. J.M. and P.V.K. drafted the manuscript, with contributions from C.J.Y. and input from other authors.

Competing Interests

P.V.K. serves on the Scientific Advisory Board to Celsius Therapeutics Inc. and Biomage Inc. P.V.K. consults National Medical Research Center for Endocrinology of the Ministry of Health of the Russian Federation. C.J.Y. is a SAB member for and hold equity in Related Sciences and ImmunAI, a consultant for and hold equity in Maze Therapeutics and a consultant for Trex Bio. C.J.Y. has received research support from Chan Zuckerberg Initiative, Chan Zuckerberg Biohub and Genentech.

Acknowledgements

We thank Anna Igolkina (St. Petersburg Technical University, Russia) for advice on the cell proportion analysis.

References

Allen, M.E., Rus, V., and Szeto, G.L. (2021). Leveraging Heterogeneity in Systemic Lupus Erythematosus for New Therapies. Trends in Molecular Medicine *27*, 152–171.

Baechler, E.C., Batliwalla, F.M., Karypis, G., Gaffney, P.M., Ortmann, W.A., Espe, K.J., Shark, K.B., Grande, W.J., Hughes, K.M., Kapur, V., et al. (2003). Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. PNAS *100*, 2610–2615.

Barkas, N., Petukhov, V., Nikolaeva, D., Lozinsky, Y., Demharter, S., Khodosevich, K., and Kharchenko, P.V. (2019). Joint analysis of heterogeneous single-cell RNA-seq dataset collections. Nature Methods *16*, 695–698.

Barkas, N., Petukhov, V., Kharchenko, P.V., and Biederstedt, Evan (2021). pagoda2: Single Cell Analysis and Differential Expression.

Bastard, P., Rosen, L.B., Zhang, Q., Michailidis, E., Hoffmann, H.-H., Zhang, Y., Dorgham, K., Philippot, Q., Rosain, J., Béziat, V., et al. (2020). Autoantibodies against type I IFNs in patients with life-threatening COVID-19. Science *370*, eabd4585.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological) *57*, 289–300.

Bennett, L., Palucka, A.K., Arce, E., Cantrell, V., Borvak, J., Banchereau, J., and Pascual, V. (2003). Interferon and Granulopoiesis Signatures in Systemic Lupus Erythematosus Blood. Journal of Experimental Medicine *197*, 711–723.

Bombardier, C., Gladman, D.D., Urowitz, M.B., Caron, D., Chang, C.H., Austin, A., Bell, A., Bloch, D.A., Corey, P.N., Decker, J.L., et al. (1992). Derivation of the sledai. A disease activity index for lupus patients. Arthritis & Rheumatism *35*, 630–640.

Bro, R. The N-way on-line course on PARAFAC and PLS.

Browaeys, R., Saelens, W., and Saeys, Y. (2020). NicheNet: modeling intercellular communication by linking ligands to target genes. Nature Methods *17*, 159–162.

Buang, N., Tapeng, L., Gray, V., Sardini, A., Whilding, C., Lightstone, L., Cairns, T.D., Pickering, M.C., Behmoaras, J., Ling, G.S., et al. (2021). Type I interferons affect the metabolic fitness of CD8+ T cells from patients with systemic lupus erythematosus. Nat Commun *12*, 1980.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol *36*, 411–420.

Cabello-Aguilar, S., Alame, M., Kon-Sun-Tack, F., Fau, C., Lacroix, M., and Colinge, J. (2020). SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics. Nucleic Acids Research *48*, e55–e55.

Catalina, M.D., Bachali, P., Geraci, N.S., Grammer, A.C., and Lipsky, P.E. (2019). Gene expression analysis delineates the potential roles of multiple interferons in systemic lupus erythematosus. Commun Biol *2*, 1–13.

Chen, S., Rivaud, P., Park, J.H., Tsou, T., Charles, E., Haliburton, J.R., Pichiorri, F., and Thomson, M. (2020). Dissecting heterogeneous cell populations across drug and disease conditions with PopAlign. PNAS *117*, 28784–28794.

Consortium, T.Gte. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science *348*, 648–660.

Corridoni, D., Antanaviciute, A., Gupta, T., Fawkner-Corbett, D., Aulicino, A., Jagielowicz, M., Parikh, K., Repapi, E., Taylor, S., Ishikawa, D., et al. (2020). Single-cell atlas of colonic CD8 + T cells in ulcerative colitis. Nature Medicine *26*, 1480–1490.

Crow, M.K., Kirou, K.A., and Wohlgemuth, J. (2003). Microarray Analysis of Interferon-regulated Genes in SLE. Autoimmunity *36*, 481–490.

Crowell, H.L., Soneson, C., Germain, P.-L., Calini, D., Collin, L., Raposo, C., Malhotra, D., and Robinson, M.D. (2020). muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. Nature Communications *11*, 6077.

Cruikshank, W.W., Berman, J.S., Theodore, A.C., Bernardo, J., and Center, D.M. (1987). Lymphokine activation of T4+ T lymphocytes and monocytes. The Journal of Immunology *138*, 3817–3823.

Davenport, E.E., Amariuta, T., Gutierrez-Arcelus, M., Slowikowski, K., Westra, H.-J., Luo, Y., Shen, C., Rao, D.A., Zhang, Y., Pearson, S., et al. (2018). Discovering in vivo cytokine-eQTL interactions from a lupus clinical trial. Genome Biology *19*, 168.

Dieudonné, Y., Gies, V., Guffroy, A., Keime, C., Bird, A.K., Liesveld, J., Barnas, J.L., Poindron, V., Douiri, N., Soulas-Sprauel, P., et al. (2019). Transitional B cells in quiescent SLE: an early checkpoint imprinted by IFN. J Autoimmun *102*, 150–158.

Dodeller, F., and Schulze-Koops, H. (2006). The p38 mitogen-activated protein kinase signaling cascade in CD4 T cells. Arthritis Res Ther *8*, 205.

Efremova, M., Vento-Tormo, M., Teichmann, S.A., and Vento-Tormo, R. (2020). CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. Nature Protocols *15*, 1484–1506.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C. (2003). Isometric Logratio Transformations for Compositional Data Analysis. Mathematical Geology *35*, 279–300.

Enocsson, H., Wetterö, J., Eloranta, M.-L., Gullstrand, B., Svanberg, C., Larsson, M., Bengtsson, A.A., Rönnblom, L., and Sjöwall, C. (2021). Comparison of Surrogate Markers of the Type I Interferon Response and Their Ability to Mirror Disease Activity in Systemic Lupus Erythematosus. Front Immunol *12*, 688753.

Fava, A., and Petri, M. (2019). Systemic Lupus Erythematosus: Diagnosis and Clinical Management. J Autoimmun *96*, 1–13.

Ferreira, R.C., Castro Dopico, X., Oliveira, J.J., Rainbow, D.B., Yang, J.H., Trzupek, D., Todd, S.A., McNeill, M., Steri, M., Orrù, V., et al. (2019). Chronic Immune Activation in Systemic Lupus Erythematosus and the Autoimmune PTPN22 Trp620 Risk Allele Drive the Expansion of FOXP3+ Regulatory T Cells and PD-1 Expression. Front Immunol *10*, 2606.

Ghodke-Puranik, Y., Imgruet, M., Dorschner, J.M., Shrestha, P., McCoy, K., Kelly, J.A., Marion, M., Guthridge, J.M., Langefeld, C.D., Harley, J.B., et al. (2020). Novel genetic associations with interferon in systemic lupus erythematosus identified by replication and fine-mapping of trait-stratified genome-wide screen. Cytokine *132*, 154631.

Giamarellos-Bourboulis, E.J., Netea, M.G., Rovina, N., Akinosoglou, K., Antoniadou, A., Antonakos, N., Damoraki, G., Gkavogianni, T., Adami, M.-E., Katsaounou, P., et al. (2020). Complex Immune Dysregulation in COVID-19 Patients with Severe Respiratory Failure. Cell Host & Microbe *27*, 992-1000.e3.

Grimbert, P., Bouguermouh, S., Baba, N., Nakajima, T., Allakhverdi, Z., Braun, D., Saito, H., Rubio, M., Delespesse, G., and Sarfati, M. (2006). Thrombospondin/CD47 Interaction: A Pathway to Generate Regulatory T Cells from Human CD4+CD25− T Cells in Response to Inflammation. The Journal of Immunology *177*, 3534–3541.

Haghverdi, L., Lun, A.T.L., Morgan, M.D., and Marioni, J.C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol *36*, 421–427.

Han, G.-M., Chen, S.-L., Shen, N., Ye, S., Bao, C.-D., and Gu, Y.-Y. (2003). Analysis of gene expression profiles in human systemic lupus erythematosus using oligonucleotide microarray. Genes & Immunity *4*, 177–186.

Hooks, J.J., Moutsopoulos, H.M., Geis, S.A., Stahl, N.I., Decker, J.L., and Notkins, A.L. (1979). Immune Interferon in the Circulation of Patients with Autoimmune Disease. New England Journal of Medicine *301*, 5–8.

Hooks, J.J., Moutsopoulos, H.M., Geis, S.A., Stahl, N.I., Decker, J.L., and Notkins, A.L. (2010). Immune Interferon in the Circulation of Patients with Autoimmune Disease (Massachusetts Medical Society).

Hsu, B.L., Harless, S.M., Lindsley, R.C., Hilbert, D.M., and Cancro, M.P. (2002). Cutting edge: BLyS enables survival of transitional and mature B cells through distinct mediators. J Immunol *168*, 5993–5996.

Iwata, Y., Wada, T., Furuichi, K., Sakai, N., Matsushima, K., Yokoyama, H., and Kobayashi, K. (2003). p38 Mitogen-Activated Protein Kinase Contributes to Autoimmune Renal Injury in MRL-Faslpr Mice. JASN *14*, 57–67.

Jin, S., Guerrero-Juarez, C.F., Zhang, L., Chang, I., Ramos, R., Kuan, C.-H., Myung, P., Plikus, M.V., and Nie, Q. (2021). Inference and analysis of cell-cell communication using CellChat. Nature Communications *12*, 1088.

Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118–127.

Juárez-Vicuña, Y., Pérez-Ramos, J., Adalid-Peralta, L., Sánchez, F., Martínez-Martínez, L.A., Ortiz-Segura, M. del C., Pichardo-Ontiveros, E., Hernández-Díazcouder, A., Amezcua-Guerra, L.M., Ramírez-Bello, J., et al. (2021). Interferon Lambda 3/4 (IFNλ3/4) rs12979860 Polymorphisms Is Not Associated With Susceptibility to Systemic Lupus Erythematosus, Although It Regulates OASL Expression in Patients With SLE. Frontiers in Genetics *12*, 785.

Kang, H.M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., Wan, E., Wong, S., Byrnes, L., Lanata, C.M., et al. (2018). Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nature Biotechnology *36*, 89–94.

Kariuki, S.N., Franek, B.S., Kumar, A.A., Arrington, J., Mikolaitis, R.A., Utset, T.O., Jolly, M., Crow, M.K., Skol, A.D., and Niewold, T.B. (2010). Trait-stratified genome-wide association study identifies novel and diverse genetic associations with serologic and cytokine phenotypes in systemic lupus erythematosus. Arthritis Res Ther *12*, R151.

King, H.W., Orban, N., Riches, J.C., Clear, A.J., Warnes, G., Teichmann, S.A., and James, L.K. (2021). Single-cell analysis of human B cell maturation predicts how antibody class switching shapes selection dynamics. Sci. Immunol. *6*, eabe6291.

Kirou, K.A., Lee, C., George, S., Louca, K., Peterson, M.G.E., and Crow, M.K. (2005). Activation of the interferon-alpha pathway identifies a subgroup of systemic lupus erythematosus patients with distinct serologic features and active disease. Arthritis Rheum *52*, 1491–1503.

Kolda, T., and Bader, B. (2009). Tensor Decompositions and Applications. SIAM Rev.

Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M.N., and Sergushichev, A. (2021). Fast gene set enrichment analysis. BioRxiv 060012.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics *9*, 559.

Li, J., Bien, J., and Wells, M.T. (2018). rTensor: An R Package for Multidimensional Array (Tensor) Unfolding, Multiplication, and Decomposition. Journal of Statistical Software *87*, 1–31.

Liu, C., Martins, A.J., Lau, W.W., Rachmaninoff, N., Chen, J., Imberti, L., Mostaghimi, D., Fink, D.L., Burbelo, P.D., Dobbs, K., et al. (2021). Time-resolved systems immunology reveals a late juncture linked to fatal COVID-19. Cell *184*, 1836-1857.e22.

Liu, M., Guo, Q., Wu, C., Sterlin, D., Goswami, S., Zhang, Y., Li, T., Bao, C., Shen, N., Fu, Q., et al. (2019). Type I interferons promote the survival and proinflammatory properties of transitional B cells in systemic lupus erythematosus patients. Cell Mol Immunol *16*, 367–379.

Liu, Z., Xue, L., Liu, Z., Huang, J., Wen, J., Hu, J., Bo, L., and Yang, R. (2016). Tumor Necrosis Factor-Like Weak Inducer of Apoptosis Accelerates the Progression of Renal Fibrosis in Lupus Nephritis by Activating SMAD and p38 MAPK in TGF-β1 Signaling Pathway. Mediators Inflamm *2016*, 8986451.

Luijten, R.K.M.A.C., Fritsch-Stork, R.D., Bijlsma, J.W.J., and Derksen, R.H.W.M. (2013). The use of glucocorticoids in Systemic Lupus Erythematosus. After 60years still more an art than science. Autoimmunity Reviews *12*, 617–628.

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.

Melé, M., Ferreira, P.G., Reverter, F., DeLuca, D.S., Monlong, J., Sammeth, M., Young, T.R., Goldmann, J.M., Pervouchine, D.D., Sullivan, T.J., et al. (2015). The human transcriptome across tissues and individuals. Science *348*, 660–665.

Nehar-Belaid, D., Hong, S., Marches, R., Chen, G., Bolisetty, M., Baisch, J., Walters, L., Punaro, M., Rossi, R.J., Chung, C.-H., et al. (2020). Mapping systemic lupus erythematosus heterogeneity at the single-cell level. Nature Immunology *21*, 1094–1106.

Nikpour, M., Dempsey, A.A., Urowitz, M.B., Gladman, D.D., and Barnes, D.A. (2008). Association of a gene expression profile from whole blood with disease activity in systemic lupus erythaematosus. Ann Rheum Dis *67*, 1069–1075.

Olwal, C.O., Nganyewo, N.N., Tapela, K., Djomkam Zune, A.L., Owoicho, O., Bediako, Y., and Duodu, S. (2021). Parallels in Sepsis and COVID-19 Conditions: Implications for Managing Severe COVID-19. Frontiers in Immunology *12*, 91.

Perez, R., Gordon, G., Subramaniam, M., Cheol Kim, M., Hartoularos, G., Sasha, T., Sun, Y., Ogorodnikov, A., Bueno, R., Lu, A., et al. Single-cell RNA-seq reveals the cell-type-specific molecular and genetic associations to lupus.

Ren, X., Wen, W., Fan, X., Hou, W., Su, B., Cai, P., Li, J., Liu, Y., Tang, F., Zhang, F., et al. (2021). COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. Cell *184*, 1895-1913.e19.

Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011). GC-Content Normalization for RNA-Seq Data. BMC Bioinformatics *12*, 480.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

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

Sheehan, B.N., and Saad, Y. (2007). Higher Order Orthogonal Iteration of Tensors (HOOI) and its Relation to PCA and GLRAM. In Proceedings of the 2007 SIAM International Conference on Data Mining, (Society for Industrial and Applied Mathematics), pp. 355–365.

Shen, Y., Ma, Y., Xie, J., Lin, L., Shi, Y., Li, X., Shen, P., Pan, X., and Ren, H. (2020). A regulatory role for CD72 expression on B cells and increased soluble CD72 in primary Sjogren's syndrome. BMC Immunology *21*, 21.

Shin, E.-C., Seifert, U., Kato, T., Rice, C.M., Feinstone, S.M., Kloetzel, P.-M., and Rehermann, B. (2006). Virus-induced type I IFN stimulates generation of immunoproteasomes at the site of infection. J Clin Invest *116*, 3006–3014.

Simon, Q., Pers, J.-O., Cornec, D., Le Pottier, L., Mageed, R.A., and Hillion, S. (2016). In-depth characterization of CD24(high)CD38(high) transitional human B cells reveals different regulatory profiles. J Allergy Clin Immunol *137*, 1577-1584.e10.

Sjöstrand, M., Johansson, A., Aqrawi, L., Olsson, T., Wahren-Herlenius, M., and Espinosa, A. (2016). The Expression of BAFF Is Controlled by IRF Transcription Factors. The Journal of Immunology *196*, 91–96.

Spinetti, T., Hirzel, C., Fux, M., Walti, L.N., Schober, P., Stueber, F., Luedi, M.M., and Schefold, J.C. (2020). Reduced Monocytic Human Leukocyte Antigen-DR Expression Indicates Immunosuppression in Critically Ill COVID-19 Patients. Anesth Analg *131*, 993–999.

Squair, J.W., Gautier, M., Kathe, C., Anderson, M.A., James, N.D., Hutson, T.H., Hudelle, R., Qaiser, T., Matson, K.J.E., Barraud, Q., et al. (2021). Confronting false discoveries in single-cell differential expression. Nat Commun *12*, 5692.

Stephenson, E., Reynolds, G., Botting, R.A., Calero-Nieto, F.J., Morgan, M.D., Tuong, Z.K., Bach, K., Sungnak, W., Worlock, K.B., Yoshida, M., et al. (2021). Single-cell multi-omics analysis of the immune response in COVID-19. Nat Med *27*, 904–916.

Stewart, A., Ng, J., Wallis, G., Tsioligka, V., Fraternali, F., and Dunn-Walters, D. (2020). Single-cell transcriptomic analyses define distinct peripheral B cell subsets and discrete development pathways. BioRxiv 2020.09.03.281527.

Stoeckius, M., Zheng, S., Houck-Loomis, B., Hao, S., Yeung, B.Z., Mauck, W.M., Smibert, P., and Satija, R. (2018). Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biol *19*, 224.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888-1902.e21.

Suen, J.-L., and Chiang, B.-L. (2012). CD4+FoxP3+ regulatory T-cells in human systemic lupus erythematosus. Journal of the Formosan Medical Association *111*, 465–470.

Teichmann, L.L., Cullen, J.L., Kashgarian, M., Dong, C., Craft, J., and Shlomchik, M.J. (2015). Local Triggering of the ICOS Coreceptor by CD11c+ Myeloid Cells Drives Organ Inflammation in Lupus. Immunity *42*, 552–565.

Tucker, L.R. (1966). Some mathematical notes on three-mode factor analysis. Psychometrika *31*, 279–311.

Unkel, S., Hannachi, A., Trendafilov, N.T., and Jolliffe, I.T. (2011). Independent Component Analysis for Three-Way Data With an Application From Atmospheric Science. JABES *16*, 319–338.

Wang, E.Y., Mao, T., Klein, J., Dai, Y., Huck, J.D., Jaycox, J.R., Liu, F., Zhou, T., Israelow, B., Wong, P., et al. (2021). Diverse functional autoantibodies in patients with COVID-19. Nature *595*, 283–288.

Wang, Y., Hicks, S.C., and Hansen, K.D. (2020). Co-expression analysis is biased by a mean-correlation relationship.

Weckerle, C.E., Franek, B.S., Kelly, J.A., Kumabe, M., Mikolaitis, R.A., Green, S.L., Utset, T.O., Jolly, M., James, J.A., Harley, J.B., et al. (2011). Network Analysis of Associations between Serum Interferon Alpha Activity, Autoantibodies, and Clinical Features in Systemic Lupus Erythematosus. Arthritis Rheum *63*, 1044–1053.

van der Wijst, M.G.P., Vazquez, S.E., Hartoularos, G.C., Bastard, P., Grant, T., Bueno, R., Lee, D.S., Greenland, J.R., Sun, Y., Perez, R., et al. (2021). Longitudinal single-cell epitope and RNA-sequencing reveals the immunological impact of type 1 interferon autoantibodies in critical COVID-19. BioRxiv.

Wikenheiser, D.J., and Stumhofer, J.S. (2016). ICOS Co-Stimulation: Friend or Foe? Front. Immunol. *0*.

Winkler, M.S., Rissiek, A., Priefler, M., Schwedhelm, E., Robbe, L., Bauer, A., Zahrte, C., Zoellner, C., Kluge, S., and Nierhaus, A. (2017). Human leucocyte antigen (HLA-DR) gene expression is reduced in sepsis and correlates with impaired TNFα response: A diagnostic tool for immunosuppression? PLoS One *12*, e0182427.

Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biology *19*, 15.

Yang, Y., Waters, J.B., Früh, K., and Peterson, P.A. (1992). Proteasomes are regulated by interferon gamma: implications for antigen processing. Proc Natl Acad Sci U S A *89*, 4928–4932.

Yung, S., and Chan, T.M. (2015). Mechanisms of Kidney Injury in Lupus Nephritis – the Role of Anti-dsDNA Antibodies. Front Immunol *6*, 475.

Yung, S., Cheung, K.F., Zhang, Q., and Chan, T.M. (2010). Anti-dsDNA Antibodies Bind to Mesangial Annexin II in Lupus Nephritis. JASN *21*, 1912–1927.

Zappia, L., Phipson, B., and Oshlack, A. (2017). Splatter: simulation of single-cell RNA sequencing data. Genome Biology *18*, 174.

Zhou, G., and Cichocki, A. (2012). Fast and Unique Tucker Decompositions via Multiway Blind Source Separation. Bulletin of the Polish Academy of Sciences: Technical Sciences *60*.

Zhou, Y., Zhang, Y., Han, J., Yang, M., Zhu, J., and Jin, T. (2020). Transitional B cells involved in autoimmunity and their impact on neuroimmunological diseases. Journal of Translational Medicine *18*, 131.