Interpretable single-cell factor decomposition using sciRED

Delaram Pouyabahar (1,2), Tallulah Andrews (7,8), Gary D. Bader (1,2,3,4,5,6,9)

1 - Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada
2 - The Donnelly Centre, University of Toronto, Toronto, Ontario, Canada
3 - Department of Computer Science, University of Toronto, Toronto, Ontario, Canada
4 - Lunenfeld-Tanenbaum Research Institute, Toronto, Ontario, Canada
5 - Princess Margaret Research Institute, University Health Network, Toronto, Ontario, Canada
6 - CIFAR Multiscale Human Program, CIFAR, Toronto, Ontario, Canada
7 - Department of Biochemistry, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Ontario, Canada
8 - Department of Computer Science, University of Western Ontario, London, Ontario, Canada
9 - Corresponding author

Abstract

Single-cell RNA sequencing (scRNA-seq) maps gene expression heterogeneity within a tissue. However, identifying biological signals in this data is challenging due to confounding technical factors, sparsity, and high dimensionality. Data factorization methods address this by separating and identifying signals in the data, such as gene expression programs, but the resulting factors must be manually interpreted. We developed Single-Cell Interpretable REsidual Decomposition (sciRED) to improve the interpretation of scRNA-seq factor analysis. sciRED removes known confounding effects, uses rotations to improve factor interpretability, maps factors to known covariates, identifies unexplained factors that may capture hidden biological phenomena and determines the genes and biological processes represented by the resulting factors. We apply sciRED to multiple scRNA-seq datasets and identify sex-specific variation in a kidney map, discern strong and weak immune stimulation signals in a PBMC dataset, reduce ambient RNA contamination in a rat liver atlas to help identify strain variation, and reveal rare cell type signatures and anatomical zonation gene programs in a healthy human liver map. These demonstrate that sciRED is useful in characterizing diverse biological signals within scRNA-seq data.

Introduction

Single-cell RNA sequencing (scRNA-seq) maps heterogeneity in gene expression in large cell populations. This heterogeneity can be attributed to various factors, both observed and hidden[1]. We can broadly categorize sources of this heterogeneity into sample-level factors, such as experimental conditions or age and weight of patients, cell-level covariates like cell type identity, cell cycle stage, and library size, or as gene-level attributes such as pathways. Each of these categories can be further classified based on their biological or technical nature and whether they are known during experimental design, or observed in or inferred based on the data. Despite their importance, identifying and interpreting factors in scRNA-seq data remains challenging due its noise, sparsity, and high dimensionality[2].

Matrix factorization (or decomposition) can identify multiple factors (signals) in a cell by gene scRNA-seq data matrix, each one capturing a unique pattern of covarying gene expression values that may represent a covariate, such as a set of genes affected by batch or a biological gene expression program[3]. Many factorization methods exist[4,5,6,7,8,9,10,11,12,13], however they generally only identify factors and leave interpretation up to the user. Some recent methods, such as scLVM[14], fscLVM[15] and

Spectra[16], can automatically interpret factors by matching them to pre-annotated gene sets (e.g. pathways), but do not consider a wider range of covariates important in single cell genomics data, such as sample-level attributes. To address this challenge, we developed Single-Cell Interpretable REsidual Decomposition (sciRED) to aid in scRNAseq factor analysis and interpretation.

sciRED enables factor discovery and interpretation in the context of known covariates and biological gene expression programs. sciRED provides an intuitive visualization of the associations between factors and covariates, along with a set of interpretability metrics for all factors. These metrics identify clear factor-covariate pairs as well as factors not matching known covariates but which are potentially interpretable. Factor-correlated genes and pathways also aid in interpretation. We apply sciRED to diverse datasets including the scMixology[17] benchmark and four biological single-cell atlases that contain known factors. We showcase its application in identifying cell identity programs and sex-specific variation in a kidney map, discerning strong and weak immune stimulation signals in a PBMC dataset, reducing ambient RNA contamination in a rat liver atlas to unveil strain-related variation, and revealing hidden biology represented by rare cell type signatures and anatomical zonation gene programs in a healthy human liver map. These demonstrate the utility of sciRED for characterizing diverse biological signals within scRNA-seq datasets.

Results

To improve the efficiency and interpretability of matrix decomposition of single-cell RNAseq data, sciRED uses four-steps to generate and characterize the resulting factors. These steps are: preprocessing data and applying matrix decomposition, evaluating factor-covariate relationships, examining unexplained factors, and determining biological interpretation of selected factors.

sciRED begins with the input cell by gene matrix and proceeds to: (1) Remove known confounding effects, factorize the residual matrix to identify additional factors not accounted for by confounding effects, and use rotations to maximize factor interpretability; (2) Automatically match factors with covariates of interest; (3) evaluate unexplained factors that may indicate hidden biological phenomena; (4) Determine the genes and biological processes represented by factors of interest (Figure 1A). In the factor discovery phase, sciRED removes user-defined unwanted technical factors, such as library size and sample or protocol, as covariates within a Poisson generalized linear model (GLM)[18]. This regresses out the covariates and produces Pearson residuals. These residuals are then factored using Principal Component Analysis (PCA) [5] with varimax rotation[19] to enhance interpretability (Figure 1B). To identify factors that explain a specific covariate, sciRED attempts to find matching labels (e.g. cell types, covariates of interest) for each factor using an ensemble classifier (see methods). Each cell is represented as a vector of factor weights, which are classified to predict covariate labels (e.g. “female” or “male” factors in the covariate “biological sex”) using four machine learning classifiers (logistic regression[20], linear classifier/area under the curve (AUC)[21,22], decision tree[23], and extreme gradient boosting (XGB)[24]). Each factor is a feature and feature selection across an ensemble of these classifiers generates factor-covariate-level association (FCA) scores, which are visualized as a heatmap. To identify high-scoring factor-covariate pairs, we use three approaches. First, we use visual inspection of the FCA heatmap. Second, we compare the association scores of each factor to the background distribution of values in the FCA table to automatically highlight significant associations (see Methods, Figure 1C). Third, we calculate a specificity measure to determine whether an explained factor captures the gene expression program related to a unique

covariate or is associated with multiple covariate levels simultaneously. Interpretation is often easier when a factor matches only one covariate level.

The third step of sciRED evaluates unexplained factors for potential interpretability using three types of metrics: separability, effect size, and homogeneity which are presented as a factor-interpretability score (FIS) table (Figures 1D, S1, see Methods). For unexplained factors that should be prioritized for follow up exploration because they may contain a hidden signal, ideally cells should be scored highly by the factor (measured by effect size), there should be two populations of cells, ones that score highly and ones that don't (bimodal distribution measured by separability) and the potential new signal is expected to be strong in cells ranked highly by that factor and other signals (e.g. known technical covariates) should be evenly distributed across this ranking (measured by homogeneity). The fourth step evaluates the factor loadings (gene lists) for biological signals. This involves manually investigating top genes and enriched pathways associated with the factor (Figure 1E).

sciRED improves factor discovery

We evaluated sciRED's factor discovery alongside eight other factor analysis methods (PCA, ICA [8], NMF [6], scVI [25], Zinbwave [9], cNMF [12], scCoGAPs [13], and Spectra[16]) on two datasets: the scMixology[17] benchmark dataset and a stimulated PBMC dataset[26], representing distinct biological and technical contexts (Figures S2-S5). The scMixology dataset includes scRNA-seq profiles from a mixture of three cancer cell lines (H2228, H1975, HCC827), sequenced across three batches with different library preparation platforms (Drop-seq, Celseq2, and 10x Genomics) representing controlled single-cell data with well-defined signals. In contrast, the stimulated PBMC dataset consists of 10x Genomics droplet-based scRNA-seq data from eight lupus patients, collected before and after a 6-hour treatment with interferon (IFN)-β, providing a real-world example of biological single-cell data.

Four metrics were used to compare performance across methods: 1) the number of entangled covariates (indicating covariates matched to multiple factors); 2) the number of factors split across multiple covariates; 3) the number of covariate levels without an associated factor; and 4) runtime. Lower values for all metrics indicate stronger performance.

sciRED outperformed other methods, with the exception of scCoGAPs, in minimizing both entangled covariates and factors distributed across multiple covariates in both datasets. All methods showed comparable results in associating covariate levels with factors, especially in capturing biological signals. However, scCoGAPs missed identifying the megakaryocyte cell type in the PBMC dataset. The runtime analysis highlights sciRED's scalability as among the top four fastest on scMixology and second fastest to PCA on the larger PBMC data set, while scCoGAPs and Zinbwave were particularly slow running over 35 hours on scMixology; and on PBMCs scCoGAPs required 139 hours, whereas Zinbwave crashed after 24 hours due to RAM limitations. All methods were run single threaded with default parameter settings on a workstation with Intel 3.0GHz Xeon E5-2687W chip and 64 GB RAM.

Notably, sciRED's runtime scales linearly with both cell and gene counts (Figure S6). This was shown using a human lung transplant dataset[27] (over 108,000 cells) and subsampling on both cell and gene counts. Runtime was tested across various cell numbers (20K, 40K, 60K, 80K, and 100K cells with genes fixed at 2000) and gene counts (500, 2000, and 5000 highly variable genes with cells fixed at 40K).

sciRED finds cell type identity programs and sex-specific processes in a human healthy kidney map

We applied sciRED to a single-cell map of healthy human kidneys (Figure 2) obtained from 19 living donors, with nearly equal contributions from both females and males (10 female, 9 male)[28]. sciRED successfully identified cell type identity programs for diverse kidney cell types (Figure 2AB, Figure S7, Table S1) as well as a factor representing sex-specific differences. For instance, factors F1 and F13 capture the identity programs of proximal tubules and endothelial cells, respectively, while F18 represents sex-specific differences (Figure 2C). Factor F18 has a high association score with male and female covariates based on the FCA heatmap, indicating it has captured sex variation. Further evaluation of the distribution of factor F18 across different cell types reveals a cell-type dependent distribution of sex-related variation, with a strong representation in the proximal tubule cell type population (Figure 2D). The FIS heatmap indicates that F18 is highly specific and shows a low homogeneity score across sex covariates, confirming that cells from male and female individuals are not uniformly distributed along F18 (Figure 2E). Consistent with the original study, pathway analysis shows an increase in processes related to aerobic metabolism (such as aerobic respiration, oxidative phosphorylation, tricarboxylic acid (TCA) cycle, and electron transport chain) in males (Figure 2FG). These findings align with the higher basal respiration and ATP-linked respiration processes in males, as functionally validated in the original study[28]. This highlights sciRED's ability to identify cell type identity signatures and sample covariate-specific variation.

sciRED identifies stimulation signals across lymphoid and myeloid cells in a stimulated PBMC dataset

We used sciRED to analyze a benchmark dataset comprising 10x Genomics droplet-based scRNA-seq PBMC data from eight lupus patients before and after a 6-hour treatment with interferon (IFN)-β[26] (Figure 3). Our sciRED analysis successfully identifies cell type identity programs and stimulation-to-control axes of variation (Figure 3A, Table S1). We identified two factors, F9 and F2, which capture stimulation signals (Figure 3B). The FIS table indicates a high bimodality (separability) score for both factors, a high effect size for F2, and high specificity for F9 (Figure 3C), illustrating the utility of the FIS table for capturing biological signals. Differential evaluation of the cell type representation within the two factors reveals a predominance of lymphoid and myeloid cells for F9 and F2, respectively (Figure 3D). In particular, F9 represents a stronger overall stimulation signal and stimulation in lymphoid populations and F2 captures a stimulation signal in myeloid populations. Examination of the cell type distribution across F9 and F2 highlights distinct clustering between the non-stimulated control and stimulated groups along both factors (Figure 3EH). Immune response-related genes and biological processes, including interferon signaling, stress or pathogen response, and cytokine signaling, are enriched in both factors (Figures 3F,G,I,J). This indicates sciRED's ability to identify both cell-type-specific gene expression activity programs, including stimulated cellular states.

sciRED alleviates ambient RNA contamination for group-based comparison

We applied sciRED to a healthy rat liver atlas mapped in Dark Agouti (DA) and Lewis (LEW) strains[29]. This atlas contains hepatocyte-derived ambient RNA contamination, a known artifact likely caused by fragile hepatocytes leaking RNA into the cell homogenate before sequencing[30]. sciRED identified factors with cell type identities, as expected, along with two factors capturing strain-associated

variation (F6 and F20) (Figure 4A-F and S8A, Table S1). The FIS table shows that factor F6 has high specificity and separability, as well as low strain homogeneity (Figure 4B). Factor F20 is the second strongest factor following F6 (Figure 4BC). F6 and F20 represent strain differences within the hepatocyte (Figure 4D) and myeloid cell types (Figure 4EF, Figures S8B and S8C), respectively. Standard differential expression analysis is not able to identify this signal due to strong ambient RNA contamination[29] (Figure 4G). For instance, four hepatocyte genes—*Fabp1*, *Tmsb4x*, *Fth1*, *Apoc1*—are among the top differentially expressed genes within the myeloid cell type of both DA and LEW strains and are estimated to be ambient RNA by SoupX[31] (Figure 4G, Table S2, see Methods). However, the top 50 myeloid strain-associated genes identified by sciRED F20 are free of such contaminants (Figure 4H, Table S1). These myeloid specific strain variations were experimentally validated in the original study. This shows the utility of sciRED in deconvolving biological signals from contamination to facilitate factor interpretation and group-based comparisons.

Exploring hidden biology in the healthy human liver atlas captured by unannotated factors

We applied sciRED to a healthy human liver atlas[32] and explored the unannotated biological signals (Figure S9, Table S1). sciRED's FCA heatmap shows that most signals correspond to liver cell type identity programs (Figure 5A). For example, F3 most strongly captures the cell identity of liver sinusoidal endothelial cells (LSECs), while F4 captures non-inflammatory macrophages. The FCA heatmap highlights nine factors (1, 10, 19, 20, 22, 26, 28, 29, 30) that are not associated with annotated cell types (Figure 5B). To discern whether these factors represent technical or biological signals, we calculated the correlation between each factor and three major technical covariates—library size, number of expressed genes, and percentage of mitochondrial gene expression—as well as cell cycle (S and G2M) phase scores (Figure 5C). Out of the nine factors, F1, F20, and F22 are correlated with technical covariates ($R > 0.45$). The remaining factors include five (F10, F19, F26, F28, and F30) that may represent unannotated biological signals, as well as F29, which shows a strong correlation with the cell cycle signature, indicating it might have captured a proliferative cell population. Evaluating the FIS table reveals that, among the six factors analyzed, all except F19 are well-mixed based on the sample covariate, suggesting a possible sample-specific effect for F19. F10 and F29 stand out with higher bimodality scores, indicating they might more effectively separate a subpopulation of cells. F26 and F28 factors exhibit a low effect size, indicating weak signals (Figure 5D). Factor F10 exhibits significant enrichment of a subpopulation of cells within the cholangiocyte cluster (Figure 5E). The top loaded genes within this factor include *MUC5A*, *MUC1*, *TFF1*, *LGALS2*, *SLPI*, *TFF2*, *TFF3*, *KRT19*, *LGALS4*, and *PIGR*. The enrichment of biological processes such as ion export, regulation of transport activity, localization, and response to ER stress strongly suggests that F10 has captured the cellular program of a rare population of mucus-producing cholangiocytes that was not identified in the original study but was reported in a more comprehensive subsequent human liver single cell map[33]. Factor F30, is enriched within a subpopulation of cells labeled as plasma cells in the original map (Figure 5F). The top genes in the loading vector include IgK+IgG+ B cell marker genes, including *IGHG4*, *IGHG1*, *IGHG3*, *IGKV1-12*, *HERPUD1*, *IGKV1-18*, *IGKC*, *SSR4*, *IGKV3-20*, and *IGKV3-21* (Figure 5F). Pathway enrichment analysis highlights biological processes such as protein folding and maturation, antibody-mediated complement activation, protein transport, and ER to cytosol transport (Figure 5F). These findings suggest that F30 may represent an antibody-secreting IgK+IgG+ B cell[33] gene expression program, not described in the original study due to its low frequency, but later captured in an expanded human liver single-cell map[33]. Factor F19 is enriched in two hepatocyte clusters, Hep1 and Hep2, which were annotated as pericentrally zoned

hepatocytes in the original study (Figure 5G). The presence of pericentral markers such as *CYP3A4*, *GLUL*, and *OAT*, along with enrichment in biological processes such as xenobiotic metabolic processes, small molecule metabolism, and lipid and fatty acid metabolism, suggests that F19 captures the anatomical pericentral signature within the hepatic lobules[30,32,34](Figure 5G). Factor F29, is correlated with cell cycle signatures and inversely enriched within a population of gd T cells (Figure 5H). The top loaded genes for F29 include cell-cycle-related regulatory genes *UBE2C*[35], *TOP2A*[36], *KPNA2*[37], *CKS2*[38], and *BIRC5*[39] and pathway analysis reveals enrichment in cell cycle and cell division, RNA splicing and processing, and nuclear division (Figure 5H). Together, these results suggest that this factor captures the cell cycle state within the gd T cell subset. We could not clearly interpret factors F26 and F28. In conclusion, sciRED is able to identify weak signals, such as rare cell types and subtle cell states, that were overlooked by standard single-cell analysis pipelines in the original study.

Discussion

We developed sciRED, a novel single cell transcriptomics data interpretation method that combines unsupervised factor analysis and supervised covariate modeling to identify biological and technical signals within single cell transcriptomics data. We introduce new metrics to assess factor interpretability to help characterize known and unknown sources of variation in the data. We also showed that regressing out known technical factors before factorization aids in data interpretation.

To test sciRED, we analyzed a series of datasets with diverse known covariates and showed that sciRED could recover these. This works well, but may miss uninterpreted signals in the original data. Although many single-cell transcriptomics simulation tools exist[9,40–45] to assist in benchmarking analysis methods, there has been little focus on factor simulation. We simulated factors to evaluate interpretability metrics, but our approach was simplistic and can be improved. sciRED uses PCA coupled with varimax rotation as a high performing matrix decomposition method, though many other factor analysis methods exist[46]. Our benchmarking results on the controlled scMixology dataset and a biologically complex PBMC dataset indicate that sciRED generally outperforms other methods in terms of combined interpretability and runtime efficiency. Tools dedicated to simulating loading and score matrices from single-cell transcriptomics, ranging from simple to complex structures, could aid in benchmarking the performance of matrix factorization methods across different scenarios. Such tools could also help answer questions such as determining the minimal sample size required to recover a factor with weak loadings. These simulation methods could draw inspiration from factor simulation techniques used in psychometric studies [47–50].

We present a suite of metrics designed to evaluate the interpretability of factors derived from a matrix factorization method within a single dataset. sciRED's modular design enables flexible integration with external data processing and factorization methods. Preprocessed data can enter directly at the factor identification stage, and previously derived factors and covariates can be imported for factor-covariate association, bypassing the initial steps if matrix factorization has already been applied using another method. Both FCA and FIS heatmaps can provide guidance for interpreting factors derived from other matrix factorization methods (Figure S10). These metrics help identify which of the K factors generated by a given factorization algorithm are likely to be interpretable. However, factor distributions can exhibit various patterns that the provided metrics may not fully capture. For unexplained factors, even one high metric value can be sufficient to highlight them as potentially interesting. In such cases,

further interpretation should rely on enriched genes and pathways. Moreover, the optimal factorization method and its specific settings—such as the number of genes included—may vary depending on the input dataset. Variables like sample complexity (including diversity of cell types, presence of rare populations, and number of covariates in the experimental design), signal-to-noise ratio (including contamination level and dropout rate), number of batches, and variation in dissociation protocols and capture technologies in multi-sample studies can influence the choice of factorization method. Developing metrics to compare and identify the optimal factor analysis method and its parameter settings to generally improve interpretability across different methods would be beneficial but remains a challenge.

In sciRED we focused on matrix factorization. However, a limitation of this approach is the assumption that cells can be represented as linear combinations of gene expression signatures, which may limit our ability to capture non-linear patterns. Alternative approaches, such as deep learning-based latent variable models like variational autoencoders (VAEs), can incorporate non-linearities and interactions between latent variables[25] in the factorization step. While the interpretability measures we developed would be applicable to VAE-derived latent spaces, non-linear patterns may inherently be more challenging to interpret. Thus, there remains a trade-off between interpretability and the ability to model complex biological phenomena effectively. The proportion of cellular programs that can be effectively captured by linear versus non-linear approaches remains uncertain for any given dataset. Enhancing the interpretability of deep learning-based models, such as VAEs, while maintaining their ability to capture complex biological phenomena is an active research area[51,52].

We only consider applying sciRED to scRNAseq data. However, it could be applied to bulk RNAseq datasets with tens to hundreds of samples. We do not recommend using the Poisson GLM step in sciRED for full-length, plate-based datasets such as Smart-seq2, as these data typically follow a different statistical distribution compared to UMI-based data [53]. However, users can apply their custom preprocessing approach and then input these data into sciRED's factor identification which would allow sciRED to be used on Smart-seq2 data. sciRED could also be extended to function with spatial and multi-omics data with multi-sample datasets, appropriate factor analysis methods and a good understanding of how technical variation affects these data. Analysis of one spatial transcriptomics dataset shows that sciRED can identify spatially distinct gene expression patterns within complex tissue structures (Figures S11 and S12). However, further refinement may be needed to enhance its handling of spatial data and other emerging data types, particularly in terms of selecting appropriate error models, and optimizing the metrics used to assess factor quality across diverse experimental technologies. We expect that as single-cell and multi-omics technologies evolve, integrating and comparing factors extracted from many datasets and various data modalities will deepen our understanding of cellular systems in healthy and diseased tissue.

Acknowledgments

This research was supported by the University of Toronto's Medicine by Design initiative, which receives funding from the Canada First Research Excellence Fund (CFREF) to GDB; by the NRNB (U.S. National Institutes of Health, grant P41 GM103504) to GDB; by the Canadian Institutes for Health Research (grant PJT 469829 to GDB) and by the University of Toronto's Data Sciences Institute Doctoral Student fellowship program.

Author contributions
Conceptualization, D.P., T.A., and G.D.B.; Methodology and software, D.P; Formal analysis, D.P., Writing – Original Draft, D.P.; Writing – Review & Editing, T.A., and G.D.B.; Supervision, G.D.B., T.A.; Funding Acquisition, G.D.B.

Declaration of interests: GDB is an advisor for Deep Genomics and is on the Scientific Advisory Board of Adela Bio. No other competing interests are declared.

Main Figure Legends

Figure 1) sciRED overview. A) sciRED comprises four main steps: factor discovery, factor interpretation, factor evaluation and biological characterization. B) In the factor discovery step, a Poisson generalized linear model is applied to the data to remove technical covariates, followed by extraction of residuals and factorization using PCA. The resulting score and loading matrices are then rotated for enhanced interpretability. The score matrix represents the projection of the original data onto the new factor space, illustrating the relationship between cells and factors. Each entry in this matrix reflects how much each cell contributes to the factors. The loading matrix contains the weights or coefficients that define the factors as linear combinations of the original genes. These weights can be used to rank genes according to their contribution to each factor, facilitating further interpretation of the factors. Models and algorithms are color-coded in blue, while input, output, and intermediate data elements are represented as white boxes. C) Factor interpretation uses an ensemble classifier to match factors with given covariates, generating a Factor-Covariate Association (FCA) table. Covariate-matched factors are identified by thresholding FCA scores based on the distribution of all FCA scores. Unannotated factors may capture novel biological processes or other covariates. D) Factor-interpretability scores (FIS) are computed for each factor. E) The top genes and enriched pathways associated with a selected factor are identified for manual interpretation.

Figure 2) Deciphering cell type identity programs and sex-specific processes in a human healthy kidney map. We applied sciRED to a single-cell healthy human kidney map obtained from 19 living donors, with similar contributions from both females and males. A) FCA heatmap with covariate levels as rows and associated factors as columns. Arrow highlights factor F18 which demonstrates a high association score with female and male covariates. B) FCA score distribution for all factors. The red line indicates the automatically defined threshold used to identify significant factor-covariate matches in the heatmap in A. C) FIS heatmap representing the interpretability scores of the selected factors. Distribution of cells over factor F1 and factor F18 colored based on D) sex and E) cell type covariates indicates the predominant representation of sex-related variation within the proximal tubule cell type. F) Pathway analysis results on the top positive (female) and negative (male) loaded genes of factor F18. G) Top 20 positively and negatively loaded genes of factor F18.

Figure 3) Identifying interferon-β stimulation signals across lymphoid and myeloid cells in a PBMC dataset. We used sciRED to analyze PBMC scRNA-seq from eight lupus patients before and after an interferon-β (IFN-β) treatment. A) FCA heatmap displaying covariate levels as rows and associated factors as columns. Arrows highlight factors F9 and F2, which capture simulation signals. B) FIS heatmap illustrating the interpretability scores of the selected factors. Red boxes highlight

factors capturing IFN-β simulation. C) Sorted factors based on FCA score values for the stimulated covariate level. F9 and F2 are the top factors associated with the stimulation covariate. D) Cell distribution over factors F9 and F2, colored based on cell type covariates. The red dashed line represents the diagonal line passing through the origin. Solid lines are the regression lines that fit each cell type. Lines with larger slopes than the diagonal represent cell types with higher association with F9, while lines with smaller slopes represent cell types more associated with F2. E) Distribution of cells over factors F9 and F1 colored based on simulation/control covariates, revealing distinct clustering between control and stimulated groups along the F9 axis. F) Pathway analysis based on the top-loaded genes of factor F9. G) Top 30 positively loaded genes of factor F9. H) Distribution of cells over factors F1 and F2 colored by simulation state covariate ("stimulated" or "control"/non-stimulated). I) Pathway analysis based on the top-loaded genes of factor F2. J) Top 30 positively loaded genes of factor F2.

Figure 4) sciRED identifies strain-based variation despite ambient RNA contamination in a rat liver map. We applied sciRED to a healthy atlas of the rat liver from two rat strains, Dark Agouti (DA) and Lewis (LEW) containing hepatocyte-derived ambient RNA contamination. A) FCA heatmap displaying covariate levels as rows and associated factors as columns. C) Factors F6 and F20 are most associated with strain variation. B) FIS heatmap illustrating the interpretability scores of the selected factors. D) Distribution of cells over factors F6 vs. F1 colored by strain and E) factors F20 vs. F1 colored by strain and F) by cell type, indicating that factor F20 captures strain variation within the myeloid population. G) Volcano plot of differential expression between strains within myeloid cells. Red dots are hepatocyte-derived ambient RNA transcripts as estimated by SoupX. Four hepatocyte genes—*Fabp1*, *Tmsb4x*, *Fth1*, and *Apoc1*—are labeled among the top differentially expressed genes within the myeloid cell type of both DA and LEW strains. These genes are among the top 50 ambient RNA transcripts derived from SoupX in all four rat liver samples (Table S2). H) Top 50 myeloid strain-associated genes identified by sciRED factor 20, free of contamination from hepatocyte-derived ambient RNA.

Figure 5) Exploring hidden biology in the healthy human liver atlas using unannotated factors. We demonstrate how sciRED facilitates the identification of hidden biology in a healthy human liver single-cell transcriptomic atlas. A) sciRED's FCA heatmap reveals signals corresponding to human liver cell type identity gene expression programs. B) Distribution of the number of matched covariate levels per factor identifies nine unannotated factors (F1, F10, F19, F20, F22, F26, F28, F29, F30) not associated with any given covariate. C) Correlation analysis between unannotated factors and technical covariates (library size, number of expressed genes, percentage of mitochondrial gene expression) and cell cycle (S and G2M) phase scores. D) FIS heatmap indicating the interpretability scores of unexplained factors uncorrelated with technical covariates. E-H) The first row shows the distribution of selected factors on the atlas tSNE plot, where each dot represents a cell, and colors indicate factor score values. The second row presents boxplots of factor scores across cell types. The third row displays the top-loaded genes, and the fourth row provides pathway analysis for factors E) F10, F) F30, G) F19, and H) F29. E) Factor 10 exhibits significant enrichment within a subpopulation of cells within the cholangiocyte cluster, suggesting the capture of a rare population of mucus-producing cholangiocytes. F) Factor 30 demonstrates positive enrichment within a subpopulation of cells labeled as plasma cells in the original map, suggesting that it captures antibody-secreting IgK+IgG+ B cells. G) Factor 19 is positively enriched in pericentrally zoned

hepatocytes (clusters Hep1 and Hep2), capturing an anatomical pericentral gene expression signature within hepatic lobules. H) Factor 29 is correlated with cell cycle signatures and inversely enriched within a population of gd T cells, suggesting capture of cell cycle state within this T cell subset.

Supplementary Figure Legends:

Supplementary Figure 1) Schematic overview of factor interpretability metrics. The four categories of factor interpretability metrics: separability, effect size, specificity, and homogeneity. Separability and effect size are label-free metrics, while specificity and homogeneity are label-dependent.

Supplementary Figure 2). **Benchmarking sciRED's factor discovery on the scMixology dataset.** We evaluated sciRED and seven other factor analysis methods (PCA, ICA, NMF, scVI, Zinbwave, cNMF, scCoGAPs) using the scMixology benchmark dataset to test factor discovery performance in single-cell data with well-defined biological and technical signals. Four evaluation metrics were used to assess performance across methods: A) the number of entangled covariates, indicating covariates matched to multiple factors; B) the number of factors distributed across multiple covariates; C) runtime; and the number of covariate levels without an associated factor. Lower values for all metrics indicate better performance. sciRED outperformed other methods, with the exception of scCoGAPs, on minimizing entangled covariates and factors split across multiple covariates. The log-scaled runtime analysis highlights sciRED's high scalability. scCoGAPs and Zinbwave showed particularly slow run time performance, requiring over 38 and 35 hours, respectively, to process the dataset. All methods successfully captured each covariate level, performing equally well on this metric.

Supplementary Figure 3) Factor-covariate association heatmaps for benchmarked methods on the scMixology dataset. Factor-covariate association tables for each method: A) sciRED, B) PCA, C) ICA, D) NMF, E) scVI, F) Zinbwave G) cNMF, H) scCoGAPs, and with initial factors set to 30 and 10. Notably, some methods exhibit higher sensitivity to the selected number of factors. For example, ICA and NMF reveal distinct splitting of cell identities for H2228 and H1975, respectively, when the factor number is set to a higher value (K=30).

Supplementary Figure 4) Benchmarking sciRED's factor discovery step on a biological dataset. We applied sciRED alongside seven other factor analysis methods (PCA, ICA, NMF, scVI, cNMF, scCoGAPs, Spectra) to a stimulated PBMC dataset to assess factor discovery performance in real single-cell data. Four metrics were defined to compare sciRED's factor discovery with other methods: A) number of entangled covariates, representing the count of covariates associated with multiple factors; B) number of factors split across multiple covariates; C) runtime; and D) number of covariate levels lacking a matching factor. Lower values in all metrics are preferable. Analysis has been performed based on the initial number of factors set to 30 for all methods. scCoGAPs algorithm internally refines the input K (number of patterns) based on the given dataset. sciRED outperformed other methods, with the exception of scCoGAPs, on minimizing entangled covariates and factors split across multiple covariates. The log-scaled runtime plot demonstrates sciRED's scalability, with it outperforming all methods except PCA in terms of runtime efficiency. Notably, scCoGAPs showed poor performance in runtime, requiring over 139 hours to process this dataset. For the number of missed factors, sciRED and cNMF each missed a single technical covariate (sciRED: sample #101;

cNMF: sample #1256), while scCoGAPs missed one biological and one technical covariates (Megakaryocytes and sample #1244).

Supplementary Figure 5) Factor-covariate association heatmaps for benchmarked methods on the stimulated PBMC dataset. Factor-covariate association tables for each method: A) sciRED, B) PCA, C) ICA, D) NMF, E) scVI, F) cNMF, G) scCoGAPs, and H) Spectra.

Supplementary Figure 6) Runtime analysis relative to the number of cells and genes. To assess sciRED's runtime scalability, we performed subsampling on the Human Lung Transplants dataset (over 108,000 cells) by varying both cell and gene counts. A) Runtime analysis across different cell counts with the number of genes fixed at 2000. The x-axis shows cell counts of 100K, 80K, 60K, 40K, and 20K. B) Runtime analysis across different numbers of highly variable genes, with the number of cells fixed at 40K. The x-axis shows gene counts of 500, 2000, and 5000. The y-axis on both scatter plots indicates runtime in seconds. Results indicate that sciRED's runtime scales linearly with both the number of cells and genes.

Supplementary Figure 7) The complete factor-covariate association table of the healthy kidney atlas. Association scores between all 30 factors and the sex and cell type covariate levels are shown. sciRED provides both the complete table and a filtered table, which includes factors with at least one association score exceeding the threshold defined over the entire distribution.

Supplementary Figure 8) Annotated healthy rat liver map. A) Association scores between all 30 factors and the strain and cell type covariate levels are shown. B) UMAP visualization of annotated healthy rat liver map, each dot represents a cell, with colors indicating cell types. C) Boxplot depicting the distribution of factor F20 across different cell types.

Supplementary Figure 9) Annotated healthy human liver map. t-SNE distribution of annotated healthy human liver map, each dot represents a cell, with colors indicating cell types.

Supplementary Figure 10) Applying sciRED's interpretation and evaluation steps to alternative factor identification methods. To assess the utility of sciRED's factor interpretation and evaluation steps on alternative factor analysis methods to PCA, we substituted sciRED's factor discovery step with NMF and applied the modified pipeline to the PBMC dataset. The resulting factors were evaluated using sciRED's standard interpretation framework. A) The FCA table shows that NMF effectively captures various cell-type identity programs and stimulation signals. However, unlike factors identified by sciRED, NMF factors often associate with multiple covariates, complicating interpretation.
B) FIS metrics were applied as in the standard pipeline. These metrics can be informative about interpretability of factors of interest. For instance, high bimodality and effect size scores for Factor 2, which is associated with megakaryocytes, highlights it as a promising factor for further evaluation.
C–D) Metrics like the number of matched covariate levels and factors per covariate are evaluated similarly to the original sciRED pipeline. E) Best-matching factors for specific covariates, such as simulation status, are identified. F) Scatter plots for Factors F5, F12, F13, and F23 illustrate how these factors capture cell-type identity, with points colored by cell type labels. G) Boxplots for these factors show score distributions across various cell types. F4 and F21 have strong associations with stimulation status, as indicated by FCA heatmap and H–I) scatter plots of cells along F1 vs. F4 and

F21, colored by stimulation status, with boxplots comparing factor scores between stimulated and control groups, as well as across cell types. Gene loading score analysis reveals that top-scoring genes in these factors include immune response-related genes, further validating their relevance.

Supplementary Figure 11) Application of sciRED to spatial transcriptomics data (subject Br8100). sciRED was applied to spatial transcriptomics data from Maynard et al. (2020), focusing on the six-layered human dorsolateral prefrontal cortex (DLPFC) to identify spatial gene expression patterns. This study used the 10x Genomics Visium platform to map the spatial organization of gene expression across the six cortical layers and white matter (WM). Each tissue block of DLPFC spans the six cortical layers and WM. Two pairs of spatial replicates were sampled from neurotypical adult donors. Each pair consisted of two directly adjacent, 10-μm serial tissue sections, with the second pair located 300 μm posterior to the first. Here, we applied sciRED to Visium samples from subject Br8100, each consisting of four total samples. Library size was adjusted using sciRED's GLM Poisson, and factor decomposition was performed on the resulting residuals. A) The FCAT heatmap shows associations between factors and cortical layers, sample IDs, and replicate covariates, with Factors 1, 2, 3, 5, and 10 capturing distinct signatures related to cortical layers and WM. B) Low correlation between factors and library size indicates these factors are minimally affected by technical variation related to read count. C) Barplot showing the number of matched factors for each covariate level. D) Boxplot distributions for Factors 1, 2, 3, 5, and 10. E-H) Projections of factor scores across each sample slide, illustrate spatial distribution. For example, Factor 1 captures a WM gene expression signature consistently across all samples, supported by its boxplot distribution and its enrichment in WM regions in each Visium slide.

Supplementary Figure 12) Application of sciRED to spatial transcriptomics data (subject Br5292). Similar to Figure S11, sciRED was applied to spatial transcriptomics samples from another subject, Br5292, demonstrating its effectiveness in capturing dorsolateral prefrontal cortex layer-specific gene expression signatures. A) The FCAT heatmap shows associations between factors, cortical layers, sample IDs, and replicate covariates. B) To assess the impact of adjusting for library size alone versus adjusting for both library size and sample IDs using a Poisson GLM, we conducted a correlation analysis, showing that both approaches yielded equivalent factors. Factors 1, 3, 5, 6, and 10 capture distinct spatial gene expression patterns, as shown by C) boxplot distributions across cortical layers and D-G) the projection of scores across the Visium slides.

Supplementary Figure 13) Enhancement in identifying cell line-specific factors through rotations. FCA tables depict A) PCA, B) sciRED, and C) Promax-rotated PCs obtained from the Pearson residual of the scMixology dataset after regressing protocol and library size covariates. D) Correlation heatmap illustrates the relationship between varimax and PCs. E) Correlation heatmap between sciRED (varimax-rotated) and promax factors.

Supplementary Figure 14) sciRED performance in the presence of strong batch effects. To assess sciRED's robustness against batch effects, we applied it to two PBMC datasets profiled with 10x Genomics single-cell 3' and 5' gene expression libraries, where batch effects between assays are substantial. sciRED was run on the combined count matrices, regressing out library and sample IDs in the Poisson GLM step. A) UMAP projection of the merged datasets colored by assay type, and B) UMAP projection colored by cell type labels, both reveal significant batch effects, with poor cell

integration observed without correction. C) FCA and D) FIS heatmaps, with examples of factors F1 and F4. E) Scatter plot showing the distribution of cells along F1 and F4, colored by cell type. F) Boxplot of F1 scores across cell type labels. G) Projection of absolute F1 values over the unintegrated UMAP, highlighting its enrichment in CD14+ monocyte clusters from both assays. H) Scatter plot of cell distribution across F1 and F4, colored by assay type. I) Boxplot of F4 scores across cell types. J) Projection of absolute F4 values over the unintegrated UMAP, showing enrichment in B cell clusters across both assays.

Supplementary Figure 15) Evaluation of sparsity effect on sciRED's decomposition results. To assess the impact of data sparsity on sciRED's performance, we varied the sparsity levels (0.01, 0.3, 0.5, 0.7, 0.9, 0.95, 0.99) in the human liver atlas dataset by progressively replacing proportions of gene expression values with zeros. The resulting factors were evaluated based on A) average variance explained across all factors, B) maximum variance explained among factors, C) percentage of matched factors, and D) percentage of matched covariates at each sparsity level. E) FCA heatmaps display factor associations for sparsity levels of 0.01, 0.5, 0.95, and 0.99 as examples. F and G show corresponding bar plots of the number of matched covariate levels per factor and the number of matched factors per covariate level, respectively.

Supplementary Figure 16) Evaluating the effect of factor number on sciRED's decomposition results based on human liver map. To investigate the impact of varying factor counts on sciRED's performance, we applied decomposition on the human liver map dataset using factor numbers (K) of 5, 10, 20, 30, and 50, comparing the outcomes. A) Number of matched covariates across different factor counts. B) Number of matched factors per covariate level. C) Correlation heatmap between factors derived from K=10 and K=30 decompositions. D and E show FCA heatmaps for K=10 and K=30, respectively. In the K=10 decomposition, three key factors—F1 (black), F3 (pink), and F7 (yellow)—were identified. For the K=30 decomposition, F1 from K=10 maps to F1, F6, F8, F25, and F27 (black box and asterisks); F3 from K=10 corresponds to F3, F13, and F16 in K=30 (pink box and asterisks); and F7 from K=10 aligns with F2, F7, F15, F18, and F21 in K=30 (yellow box and asterisks). These results indicate that limiting factor numbers may cause individual factors to aggregate multiple gene expression programs, reducing interpretability.

Supplementary Figure 17) Evaluating the effect of factor number on sciRED's decomposition results using the stimulated PBMC dataset. We analyzed the impact of varying factor counts (K) on the decomposition results for the PBMC dataset by applying factor numbers of 5, 10, 20, 30, and 50. A) Number of matched covariates for each factor count. B) Number of matched factors per covariate level. Results were consistent with those from the human liver map: as the factor count increased, each factor tended to align with fewer covariates, and a larger proportion of factors remained unmatched at higher K values.

Supplementary Figure 18) Performance evaluation and benchmarking of sciRED. A) Permutation tests were used to assess the significance of each covariate-factor association metric (see methods). B) Comparison of AUC, K-Nearest Neighbors (KNN), logistic regression, decision tree, random forest, and XGB using permutation tests on the scMixology benchmark dataset. We expect the unshuffled (baseline) scores to be higher than the shuffled association scores. KNN's poor performance led to its exclusion from the sciRED ensemble model. The importance scores of each model were min-max

scaled for comparison. C) Run time evaluation resulted in the exclusion of the random forest method due to inferior scalability performance. D-G) Benchmark analysis illustrates sciRED's superior performance relative to single classifiers. We assessed the average number of significant matched factors per covariate level for the D) human kidney map, E) human liver map, F) scMixology dataset and G) stimulated PBMC dataset at a significance level of p-value=0.05. The dashed line denotes the value of one factor-covariate association. Model distributions closer to a mean of one and lower variance suggest improved specificity and stability. (ns: p > 0.05, *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001)

Supplementary Figure 19) Optimizing ensemble design through permutation-based comparison of scaling and mean calculation methods. We compared three scaling methods (standardization, min-max scaling, and rank-based) and two mean calculation approaches (arithmetic and geometric) to design an optimal ensemble. sciRED uses standardization scaling and arithmetic mean calculation, which is the optimal combination based on analysis in this figure (panels A, B are for results based on the scMixology dataset and panels C, D are for results based on the healthy human liver dataset). We used the average number of significant associated factors per covariate (p-value=0.05) along the Gini coefficient[54] (see methods) for comparing ensemble designs. Ideally, the average number of significant associated factors per covariate would be close to one, and the Gini coefficients of the shuffled datasets are expected to be less than for the unshuffled data. A) Average number of significant associated factors per covariate (p-value=0.05). The dashed line indicates the value of one factor-covariate association, while the grey area represents the approximate acceptable range. Distributions closer to a mean of one indicate improved specificity. B) Distribution of Gini coefficient calculated for FCA tables of shuffled datasets for each ensemble table. Red dots represent the Gini coefficient of the unshuffled dataset's FCA table, expected to be higher than for the shuffled data (see methods). C) Box plot representing the average number of significant associated factors per covariate (p-value=0.05) for the healthy human liver dataset based on various residual types. Deviance, response and Pearson residuals show comparable performance with Pearson showing lower variance for sciRED (standardization, arithmetic mean). D) Boxplots indicating the distribution of the Gini coefficient calculated for FCA tables of shuffled datasets for each ensemble model and residual type. Deviance, response, and Pearson residuals show comparable results. This analysis reveals that sciRED exhibits superior performance compared to geometric-standard, geometric-rank, and arithmetic-rank methods, as evidenced by the Gini measure. Furthermore, it outperforms ensembles utilizing min-max (which tend to exhibit high false negatives) and rank-scaled methods (associated with high false positives) when considering the average number of significant factors per covariate. The ensemble model was constructed by standardizing the importance scores of AUC, logistic, XGB, and decision tree, followed by arithmetic mean calculation.

Supplementary Figure 20) Impact of residual choice on factor interpretability. We investigated how various residual-based normalization methods affect downstream analysis, as explored by several studies[11,55,56]. Specifically, we examined the impact of different residual types (response, Pearson, and deviance) on factor interpretability performance using the healthy human liver atlas. We employed a permutation test to evaluate the significance of each covariate-factor association based on empirical p-value calculation. Using this approach, we calculated the average number of significant associated factors per covariate (p-value =0.05). Our permutation analysis revealed consistent outcomes across all residual types for most models, with the exception of the response residual-based

logistic regression model, which exhibited less stable results. This analysis underscores the robustness of sciRED's performance across different residual types. Consequently, we chose the Pearson residual as the default approach, given its widespread adoption in the field. The dashed line indicates the value of one factor-covariate association, and the grey area represents the approximate acceptable range. E) Similar to Figure D, a permutation analysis was performed on the healthy human liver atlas to evaluate the impact of residual types. Investigation of response, Pearson, and deviance residuals led to consistent performance of models considering known covariates in the scMixology dataset, with sciRED outperforming single classifiers.

Supplementary Figure 21) Evaluating factor interpretability metrics through factor simulation. Factors were simulated to evaluate the effectiveness of the proposed factor interpretability metrics.
A) Simulation assumes that factors are generated from a Gaussian mixture distribution, where each Gaussian represents cells associated with a covariate level. Correlation between the overlap of the two Gaussian distributions and the metric value was used to assess interpretability. B) Factors with greater overlap values would exhibit lower separability and specificity scores (negative correlation), along with higher homogeneity values (positive correlation), and vice versa. C) Distribution of correlation scores for each metric across various categories. The ideal correlation value is represented by a dashed red line for each interpretability category.

Supplementary Tables:
Table S1. Factor loading tables for all biological datasets.
Table S2. SoupX ambient RNA contamination estimates for each rat liver sample.

Methods

sciRED factor discovery framework

We model the read counts for n cells and g genes (i.e., $Y_{n\times g}$) as a combination of annotated and unannotated factors as follows:

$$Y_{n\times g} = C_{n\times p}\beta_{p\times g} + F_{n\times f}A_{f\times g} + U_{n\times u}H_{u\times g} + \varepsilon_{n\times g}$$

Where:
$Y_{n\times g}$ represents the observed data matrix, typically with dimensions $n \times g$ (where n is the number of samples or cells and g is the number of variables or genes)
$C_{n\times p}$ is the matrix of technical covariates with dimensions $n \times p$ (where p is the number of covariates), such as library size, batch ID, cell cycle stage
$\beta_{p\times g}$ represents the coefficient matrix for technical covariates with dimensions $p \times g$
$F_{n\times f}$ is the matrix of annotated factors with dimensions $n \times f$ (where f is the number of factors)
$A_{f\times g}$ is the matrix of loadings for annotated factors with dimensions $f \times g$
$U_{n\times u}$ is the matrix of unannotated factors with dimensions $n \times u$ (where u is the number of unannotated factors)
$H_{u\times g}$ represents the matrix of loadings for unannotated factors with dimensions $u \times g$
$\varepsilon_{n\times g}$ is the error term, with dimensions $n \times g$

Annotated factors are those matched with known (given) covariates as indicated in the factor-covariate association (FCA) table. The solution to the above equation is reached through a two-step process.

1. Derivation of residuals

The first step aims to remove the effect of technical covariates from the data.

Input counts, represented by the matrix $Y_{n\times g}$ are modeled as a Poisson generalized linear model (GLM) to account for the matrix's distributional properties[11,55,57]. Technical covariates $C_{n\times p}$ are incorporated into the model to capture their effects on the count data. The statsmodels package[58] (version 0.11.0) is used for GLM implementation.

Pearson residuals from this model are computed as the difference between observed and predicted counts divided by the square root of the predicted counts:

$$r_i = \frac{y_i - \hat{y}_i}{\sqrt{\mathrm{VAR}(\hat{y}_i)}}$$

where VAR is the variance. For Poisson GLM, $\mathrm{VAR}(\hat{y}_i) = \hat{y}_i$

The Pearson residuals are used for subsequent matrix decomposition.

Two additional residual types were evaluated, but not used in sciRED. Response residuals represent the difference between the observed count and the predicted mean count for each observation.

Response Residuals: $resid_i = y_i - \hat{y}_i$

Deviance residuals represent the individual contributions of samples to the deviance D, calculated as the signed square roots of the unit deviances. Deviance Residuals:

$$d_i = \mathrm{sign}(y_i - \hat{y}_i) \times \sqrt{2 \times (y_i \log(y_i/\hat{y}_i) - (y_i - \hat{y}_i))}$$

Pearson residuals are used for sciRED because they are established in other single cell analysis methods[55,56]. However, sciRED results are robust to the choice of residual (among Pearson, response, deviance) (See method section "Classifier ensemble").

2. Residual decomposition

After obtaining Pearson residuals from the count data, a matrix factorization technique is employed to uncover underlying patterns, including the annotated and unannotated factors. PCA factors are calculated using Singular Value Decomposition (SVD)[59] as implemented in scikit-learn[60] (version 0.22.1). Following factor decomposition, we apply varimax rotation to enhance the interpretability of the principal axes in the feature space. To achieve this we reimplemented the optimized estimation procedure as described below.

Rotation types

Factor analysis typically comprises two sequential stages. Initially, loadings are computed to best approximate the observed variances within the data. However, these initial loadings may lack interpretability, thus we apply rotation to generate a revised set that is more easily interpretable. There

are two primary rotation types in factor analysis: orthogonal and oblique rotations. Orthogonal rotation, such as varimax, seeks to produce orthogonal factors with zero correlations. Intuitively, varimax rotation aims to identify factors associated with a limited number of variables, thereby it promotes the detection of distinct factors rather than those affecting all variables evenly. Mathematically, interpretability is achieved by maximizing the variance of the squared loadings along each principal component axis. Varimax rotation seeks to maximize the Kaiser criterion:

$$v = \sum_j v_j = \sum_j \left\{ \left[g \sum_i \left(l_{ij}^2 / h_i^2 \right)^2 - \left(\sum_i l_{ij}^2 / h_i^2 \right)^2 \right] / g^2 \right\}$$

Where v is the Kaiser criterion and l_{ij} is the loading value of ith gene and jth factor, and g is the total number of genes. The communality h_i is calculated as the squared sum of the loading values for each gene.

$$h_i^2 = \sum_j l_{ij}^2$$

To explicitly specify the rotation matrix, we can reformulate the Kaiser criterion using the following notation. Let L be a $g \times k$ loading matrix (eigenvectors), and R denote a rotation matrix such that $R^T R = I_k$, where I_k is the $k \times k$ identity matrix. Additionally, let R_{ij} represent the scalar element in the ith row and jth column of matrix R. Varimax rotations can now be described as follows:

$$\mathbf{R}_{\text{varimax}} = \arg\max_{\mathbf{R}} \left(\sum_{j=1}^{k} \sum_{i=1}^{g} (\mathbf{LR})_{ij}^4 - \frac{1}{g} \sum_{j=1}^{k} \left(\sum_{i=1}^{g} (\mathbf{LR})_{ij}^2 \right)^2 \right)$$

where R_{varimax} denotes the resulting rotation matrix.

The rotation matrix is computed using an iterative method relying on SVD to achieve sparsity in the loadings. Subsequently, the rotated loadings and score matrices are derived by multiplying the original loadings and score matrices with the rotation matrix, respectively. The optimization algorithm is elaborated on in detail in Appendix A of Stegmann *et al*.[61]. We re-implemented the base R varimax rotation function in Python.

Oblique rotations, such as promax, allow factors to be correlated, thereby relaxing the orthogonality assumption[62]. This flexibility can be beneficial when factors are expected to be correlated within the underlying structure of the data. Promax initially applies the varimax method to generate a set of orthogonal results. Subsequently, it constructs an ideal oblique solution to exaggerate the orthogonal rotated gene-loading matrix. Finally, it rotates the orthogonal results to achieve a least squares fit with this ideal solution.

We define a pattern matrix $P = (p_{ij})$, as the following ($k > 1$):

$$p_{ij} = \frac{\left| l_{ij}^{k+1} \right|}{l_{ij}}$$

Each element of P matrix is the kth power (typically 3rd power) of the corresponding element in the row-column normalized varimax loading matrix. Next, the least squares fit of the orthogonal matrix of factor loadings to the pattern matrix is computed.

$$R_{\text{Promax}} = \underset{R}{\operatorname{argmin}} \|P - L_{\text{rot}} R\|_2$$
$$R_{\text{Promax}} = \left(L'_{\text{rot}} L_{\text{rot}}\right)^{-1} L'_{\text{rot}} P$$

where R_{Promax} is the unnormalized rotation matrix, L_{rot} is the varimax rotated loadings, and P is the pattern matrix defined above.The columns of R_{Promax} are normalized such that their sums of squares are equal to unity. We reimplemented the base R promax function in Python.

To evaluate the impact of factor rotations, we applied PCA, sciRED (varimax-based[19]), and promax-rotated[63] PCA to the Pearson residual of the scMixology dataset after regressing out protocol and library size. The factor-covariate association scores reveal a high correlation between sciRED and Promax. Both methods enhance specificity and achieve one-to-one association between factors and cell line covariates, outperforming unrotated PCA (Figure S13). We used varimax as the default rotation method for sciRED, as it has been established to perform well for interpretable factor analysis [64,65].

Factor identification benchmarking

For both scMixology and PBMC datasets, PCA and ICA were applied to normalized (library-regressed) data, while NMF, scVI, Zinbwave, cNMF, and scCoGAPS were run on filtered raw counts. All methods used the top 2,000 highly variable genes, following sciRED's standard pipeline. The number of factors (K) was set to 30 across all methods; for additional comparison, K was also set to 10 for the scMixology dataset.

Spectra was applied exclusively to the PBMC dataset, as it relies on the Cytopus knowledge base, which contains gene sets tailored for standard scRNA-seq data but lacks coverage for cancer cell line-specific gene sets. For PBMC, cell type labels were standardized to CD4-T, mono, cDC, NK, CD8-T, and B, with an "all-cells" category used for cell types without specific annotations. The number of factors per cell type was set to five to approximate the total number of factors used in the other methods.

All methods were run in single-threaded mode with default parameters on a workstation with Intel 3.0GHz Xeon E5-2687W chip and 64 GB RAM.

Data pre-processing and handling of batch effects with sciRED

sciRED takes raw count (not batch corrected) data as input, removing cells and genes with zero total read count and retaining only the top highly variable genes, typically the top 2000 as in standard scRNA-seq workflows. While using all genes is an option, the first step—fitting the Poisson GLM—becomes more time-consuming with larger gene sets, thus a smaller set of highly variable genes is used. The minimum required covariate is library size (total read count per cell), which serves as an equivalent normalization step.

Additional covariates can be incorporated into the GLM model based on specific analysis goals; however, technical covariates should be removed with caution to prevent the unintentional exclusion of biologically relevant signals. For instance, in analyses focused on sex-specific variations, adjusting for sample IDs as covariates can inadvertently eliminate sex-associated signals if sample and sex effects are confounded.

In multi-sample analyses, we recommend merging all samples without batch correction. Batch correction is generally unnecessary with sciRED, as our analyses demonstrate robust batch effect handling without directly correcting the data matrix or latent embedding. To evaluate batch effect handling, sciRED was tested on two PBMC datasets profiled using 10x Genomics single-cell 3' and 5' gene expression libraries with strong batch effects (Figure S14). sciRED was applied to the combined count matrices, with library and sample IDs regressed as covariates in the Poisson GLM step. The UMAP projection (Figures S14AB) reveals batch-related clustering without correction; however, sciRED analysis (Figures S14CD) shows well-integrated cell type identities, even across assay types. For instance, Factors F1 and F4, which identify CD14+ monocytes and B cells respectively, are well integrated across both assays. The distributions of cells over F1 and F4, colored by cell type and assay, demonstrate effective batch integration, and the corresponding box plots and UMAP enrichment patterns further confirm these results (Figures S14E-J).

Impact of data sparsity on factor decomposition performance

To assess the impact of data sparsity on sciRED's decomposition, we we systematically varied the sparsity levels (0.01, 0.3, 0.5, 0.7, 0.9, 0.95, 0.99) in the human liver atlas dataset by incrementally replacing randomly selected gene expression values with zeros (Figure S15). For each level, we evaluated the factors based on average and maximum variance explained, the percentage of matched factors, and matched covariates. Our results indicate that the average variance explained by all factors remains largely robust across different sparsity levels, likely due to sciRED's rotation step, as this trend is not observed for PCA factors. The maximum variance explained by factors is also more stable with sciRED's rotated factors compared to PCA as sparsity increases. As expected, the percentage of matched factors decreases with rising sparsity, with a marked drop around a sparsity threshold of ~70%. At lower sparsity (0.01 and 0.5), the FCA heatmaps display a balanced distribution of matched covariates per factor, while at higher sparsity (0.95 and 0.99), most covariates align with only a few factors, reducing interpretability. However, the percentage of matched covariates remains stable, as most covariates continue to align with at least one factor.

Impact of factor number (K) on decomposition results

To evaluate the impact of factor count (K) on sciRED's decomposition results, we tested various factor numbers—5, 10, 20, 30, and 50—on the human liver map (Figure S16). At lower factor counts, such as K=5, each factor showed higher covariate associations (e.g., Factor 1 was matched to nine covariates). As K increased, each factor aligned with fewer covariates, and the percentage of matched factors decreased. At K=50, 54% of factors had no covariate matches. Correlating factors between the K=10 and K=30 decompositions revealed that individual factors in K=10 (e.g., F1, F3, and F7) mapped to multiple factors in K=30, suggesting that low factor limits may lead to each factor capturing multiple gene expression programs, complicating interpretation. Repeating this analysis on the PBMC dataset

yielded similar results: as the number of factors increased, each factor aligned with fewer covariates, and a higher proportion of factors remained unmatched at larger K values (Figure S17).

The ideal factor count varies by dataset and is affected by biological and technical signal diversity. While a Scree plot is commonly used in PCA to determine factor counts, sciRED's rotation step changes the ordering by variance explained, making this approach less effective. Empirically, a moderately high K (e.g., 30) has yielded interpretable results across datasets. We recommend starting with a larger K and excluding factors with no covariate matches or low factor importance scores (FIS).

Feature importance calculation

We evaluated a range of classifiers for inclusion in the sciRED ensemble classifier. Each classifier is trained on individual levels of each covariate separately (e.g. "female" and "male" for a "biological sex" covariate). Each classifier uses a different approach to estimate feature importance:

1) Logistic regression: feature importance is the magnitude of the coefficient for each factor, which represents the change in the log-odds of belonging to a covariate level per unit of factor weight.
2) Decision trees, random forest[66] and extreme gradient boosting: feature importance scores represent the decrease in covariate mixing (e.g. Gini impurity or entropy) when the feature is used within a tree averaged across all trees.
3) K-nearest neighbor[67] (KNN): feature importance is estimated as the decrease in predictive accuracy when the values for that feature are randomly permuted. This value is calculated as the average across five permutations for each factor based on the default scikit-learn package implementation.
4) Linear classifier (AUC): feature importance for a linear classifier (i.e. fixed threshold) is calculated as the area under (AUC) the receiver-operating characteristic (ROC) curve. The AUC for one-dimensional data is equivalent to the Wilcoxon or Mann-Whitney U test statistic with the relation:

$$\mathrm{AUC} = \frac{U}{n_0 \times n_1}$$

Where U is the Mann-Whitney U statistic, and n_0 and n_1 are the sample sizes of the two groups being compared. Mann-Whitney U test is a non-parametric test used to assess whether two independent samples are selected from populations having the same mean rank. Here, samples are defined as factor scores for the target group (cells labeled with the covariate level of interest) and the non-target group. In the context of feature importance, a higher AUC value indicates that the factor is better at separating the classes, while a lower AUC value suggests less discriminatory power. The scikit-learn package is used to implement decision tree, random forest, logistic regression and KNN classifiers with default parameters. XGBoost (version 1.5.0) and scipy[68] (version 1.4.1) packages are used for XGB and the Mann-Whitney U test, respectively.

Classifier ensemble

We optimized the sciRED classifier ensemble by evaluating different classifier combinations on four independent datasets: a healthy human kidney map, a healthy human liver map, a PBMC atlas, and the scMixology benchmark dataset (Figure S18). For each experiment, we randomly shuffled the covariate labels to generate a null distribution of classifier association scores and calculated the

average number of significant associated factors ($p < 0.05$) per covariate level (Figure S18A). We defined a one-to-one association between factors and covariates as the optimally interpretable result. This analysis shows that the sciRED classifier (ensemble of logistic regression, linear classifier/area under the curve (AUC), decision tree and extreme gradient boosting (XGB)) outperforms or matches the performance of the individual classifiers, depending on the data set. Initially, six classifiers—AUC, K-Nearest Neighbors (KNN), logistic regression, decision tree, random forest, and XGB—were compared on the scMixology benchmark dataset. Due to KNN's poor classification performance and the random forest's inferior scalability, they were excluded from the sciRED ensemble model (Figures S18BC). Benchmarking on the four independent datasets described above demonstrates sciRED's superior performance relative to single classifiers (Figures 18D-G).

To optimize the ensemble score, we also tested all combinations of three different scaling methods and two mean calculations (Figure S19). Specifically, we considered standardization, min-max scaling, or rank-based scaling combined with arithmetic or geometric means[69]. By comparing scores for the real scMixology data vs. permuted data labels, we identified standardization followed by arithmetic mean calculation as the optimal scoring method (Figure S19).

The standardization, min-max scaling, and rank-based scaling are defined as follows:
Given a set of feature importance scores for each factor i :$x = \{x_1, x_2, \ldots, x_n\}$, and their ascending order of ranks $r = \{r_1, r_2, \ldots, r_n\}$, where n is the total number of features (factors):

$$\text{standardized}_i = \frac{x_i - \text{mean}(x)}{\text{std}(x)}$$

$$\text{min-max scaled}_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

$$\text{rank-scaled}_i = \frac{r_i}{n}$$

Rank scaling generates a value between 0 and 1 to each data point based on its rank within the dataset. Lower values in the original dataset will have lower rank-based scaled values, while higher values will have higher rank-based scaled values.

Arithmetic and geometric means across classifiers are calculated as follows:

$$\text{Arithmetic mean} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Where n is the number of values.

$$\text{Geometric mean} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}$$

The geometric mean is the nth root of the product of a set of values.

The effect of the choice of residual (Pearson, response, deviance) on sciRED performance was also evaluated, indicating that sciRED results are robust to the choice of residual (Figure S20).

Determining significant factor-covariate associations

FCA scores are binarized into significant and non-significant associations based on a threshold. This threshold is automatically determined using Otsu thresholding[70]. Otsu thresholding iterates through all potential threshold values and computes a separability measure in each iteration. The threshold value that maximizes this separability measure is chosen as the optimal threshold to partition the FCA score distributions into significant and non-significant associations. scikit-image package[71] (version 0.23.2) is used for implementation.

Benchmarking covariate-factor association metrics based on permutation results

We used a permutation test to evaluate the significance of each covariate-factor association. This process entailed randomly shuffling cell covariate labels 500 times and recalculating factor-covariate association scores for each permutation to create an empirical null distribution. A shuffled dataset should result in lower-scaled factor-covariate association scores compared to the original dataset, given that the cell labels are randomized after each permutation.

Empirical p-values were calculated as the number of permuted FCA scores are as or more extreme than the observed association value divided by the number of permutations. The number of significant associations for each covariate was determined using p-value < 0.05. We expect one significant factor association per covariate level in the ideal case. High values of this metric likely indicate false positive associations (low specificity) and zero values highlight false negative results (low sensitivity).

To evaluate model performance without relying on predefined thresholds (such as p-value=0.05), we employed the Gini index, an inequality measure ranging from 0 to 1. A Gini index of 0 signifies perfect equality, where all values are identical, while a score of 1 indicates perfect inequality, with one value dominating the distribution. In our context, we aimed to assess the extent of association between factors and covariates across various levels of pre- and post-permutation. Following label shuffling, we expect the factor-covariate association scores to exhibit uniformity across all covariate levels (Gini score closer to 0). In cases of non-random baseline labels, sparse associations between factors and covariates would result in a Gini score closer to 1. To calculate the Gini index for a given FCA matrix, Gini was calculated for each covariate level separately and the average Gini index across all covariate levels is provided.

$$G_k = \frac{1}{2n^2\bar{x}} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|$$

Where G_k is the Gini coefficient for a covariate level k, n is the number of factors, and $\bar{x}$ are the mean FCA scores for a given covariate level. We then calculate the global Gini index as:

$$G = \frac{1}{K} \sum_{i=1}^{K} G_k$$

Where K is the total number of covariate levels.

Factor evaluation

We defined four categories of metrics: separability, effect size, specificity, and homogeneity to measure factor interpretability. Two are label-free (separability, effect size) and the other two are label-dependent (specificity, and homogeneity). Metrics for each category are organized into a table of factor interpretability scores (FIS), with metrics arranged as rows and factors as columns.

Subsequently, the FIS undergoes a row-wise scaling and is visualized as a heatmap to support comparison of interpretability scores between factors.
sciRED uses the Silhouette score and bimodality index to assess separability. Homogeneity is measured using the arithmetic mean of the average scaled variance (ASV) table. The Simpson index is employed to evaluate factor specificity, while factor variance serves as an effect size measure (Figure S10, See method section "Correlation between interpretability metrics and overlap values").

Separability (bimodality)
We use bimodality scores as an indicator for measuring the separability of factors. We use existing metrics developed for identifying genes with bimodal expression distributions, which are typically used to identify genes with prognostic value in discriminating patient subgroups. We selected metrics in accordance with Hellwig et al.'s comparative assessment of these scores based on survival times in a breast cancer dataset[72].

Cluster-based bimodality measures
Cluster-based methods group observations into clusters then calculate various statistics to measure the degree of distinctiveness of these clusters. Clusters were generated using a standard k-means [73] algorithm (k=2) implemented using the scikit-learn[60] package. The following cluster-based bimodality measures are included in sciRED:

- Variance Ratio Score (VRS) also known as Calinski–Harabasz index (CHI)[74] assesses the proportion of variance reduction when splitting the data into two clusters. We can decompose the total sum of squares (TSS) to between-cluster sum of squares (BSS) and within-cluster sum of squares (WSS). VRS is then defined as the ratio of BSS and WSS.
- Weighted Variance Ratio score (WVRS) is similar to VRS but measures the variance reduction independent of the cluster sample sizes. In both cases, higher values indicate better separation between clusters and reflect bimodality.
- Silhouette score[75] measures the cohesion and separation of clusters, ranging from -1 to 1, where higher values indicate better-defined clusters. Silhouette is calculated based on mean intra-cluster distance (ICD) and the mean nearest-cluster distance (NCD). For each data point Silhouette is then given as (NCD - ICD)/max(ICD, NCD).
- Davies-Bouldin index[76] computes the average similarity between each cluster and its most similar cluster. It measures similarity as the ratio of within-cluster distances to between-cluster distances. The minimum value is zero with lower values indicating better cluster separation. To ensure consistency with other metrics, we scale the inverse Davies-Bouldin Index.

Calinski–Harabasz, Silhouette and Davies-Bouldin Indices were implemented using scikit-learn.

Bimodality index
The Bimodality Index[77] is another bimodality metric which has initially been introduced to rank bimodal gene expression signatures within cancer gene expression datasets. We adopt the bimodality index to assess the bimodal nature of factor scores by modeling them as a mixture of two normal distributions using a Gaussian Mixture Model (GMM). Let μ_1 and μ_2 represent the means of the two normal distributions, and σ the standard deviation. Given an equal variance assumption, the standardized distance δ between the distributions is calculated by:

$$\delta = \frac{|\mu_1 - \mu_2|}{\sigma}$$

Given GMM estimated parameters, we can calculate the proportion of observations in the first component π . We then compute the bimodality index (BI) as:

$$\mathrm{BI} = \sqrt{\pi(1-\pi)} \times \delta$$

A higher BI indicates a stronger bimodal distribution, aiding in the identification of bimodal factor signatures. Scikit-learn was used to fit GMMs.

Dip score

Another bimodality evaluation metric is Hartigan's dip test[78], a statistical measure used to assess deviations from unimodality in distributions. It evaluates multimodality by comparing the maximum difference between the empirical distribution function and the unimodal distribution function that minimizes this difference. The dip score was computed for each factor, indicating the degree of bimodality present in the factor distribution. Higher dip scores suggest stronger evidence of bimodality, while lower scores indicate unimodality. Implementation was based on the diptest package.

Effect size

The variance of factor scores across cells was employed as a measure of effect size consistency.

Specificity

We assessed the Simpson diversity index and Shannon entropy as measures of specificity.

Simpson diversity Index

The Simpson diversity index is a measure commonly used in ecology to quantify the diversity or evenness of species within a community[79]. It assesses the probability of encountering different species within a community and how evenly distributed these species are. We adopted the Simpson diversity index to measure specificity for individual factors within the context of FCA scores. By applying the Simpson index to each vector of FCA scores, we can evaluate the extent to which a factor uniquely characterizes a particular covariate level.

Mathematically, the Simpson diversity index D is expressed as

$$D = \sum_{i=1}^{n} p_i^2$$

where p_i denotes the scaled score of factor-covariate level association (probability of a factor being chosen for a given covariate level i), and n represents the total number of covariate levels. Here, the Simpson diversity index ranges between 0 and 1, where 0 indicates low factor specificity (maximum diversity, all factor association scores are equally distributed) and 1 indicates greater specificity (minimum diversity) of a factor towards a particular covariate level.

Shannon diversity index

We applied Shannon entropy to measure the specificity of individual factors within FCA scores. By calculating Shannon entropy for each vector of FCA scores, we assess how uniquely a factor characterizes a particular covariate level.

Mathematically, Shannon entropy is defined as:
$H = -\sum_{i=1}^{N} p_i \log(p_i)$

where p_i is the probability of a factor being associated with a given covariate level i, and N is the total number of covariate levels. Shannon diversity index ranges from 0 (high factor specificity) to $\log(N)$ (low factor specificity).

Homogeneity
To assess the homogeneity or even distribution of factor scores across different levels of a covariate, we compute the scaled variance for each covariate level. This metric quantifies the proportion of variance in the factor scores observed at a specific covariate level (L) relative to the total variance across all covariate levels[80]. Thus, the scaled variance SV for a factor x and a particular covariate level L is computed as

$$SV = \frac{\mathrm{Var}(x_L)}{\mathrm{Var}(x)}$$

Here, $\mathrm{Var}(x_L)$ denotes the variance of the factor scores x corresponding to the covariate level L, and $\mathrm{Var}(x)$ represents the total variance of the factor scores across all cells.

To establish a unified metric across all levels of a single covariate (e.g., Batch1, Batch2, etc., representing levels of the covariate "Batch"), we adopt different approaches based on the covariate's number of unique levels. We compute both the geometric and arithmetic means of the scaled variances (SV) of all covariate levels for each factor. The arithmetic mean of SV values was chosen based on factor simulation results showing it performed best (explained below).

Evaluating factor interpretability metrics using simulation
Simulating mixture Gaussian distribution
Factors were simulated under the assumption of being generated from a mixture of Gaussian distributions. Each Gaussian distribution represents the factor scores of the cells belonging to one covariate level. This simulation process was implemented using the following mathematical formulation:

Let X be a random variable representing the simulated factors. We generated n samples (cells) from a mixture of K (K=2) normal distributions, each characterized by its mean μ_k, standard deviation σ_k, and proportion p_k. The mixture distribution X follows the distribution:
$\sum_{k=1}^{K} p_k \cdot \mathcal{N}(\mu_k, \sigma_k^2)$

where $\mathcal{N}(\mu_k, \sigma_k^2)$ denotes the probability density function (PDF) of the k-th normal distribution.

Given this assumption, 10 factors for 10000 cells were simulated for 100 rounds. p_k was set to 0.5 for simplicity. μ and σ were set by random sampling from uniform distributions (parameters: $\sigma_{\min}$=0.5, $\sigma_{\max}$=1, $\mu_{\min}$=0, $\mu_{\max}$=4).

Correlation between interpretability metrics and overlap values

We assessed the efficacy of the proposed factor interpretability metrics by examining their correlation with the overlap between the two Gaussian distributions representing each factor. We anticipated that factors exhibiting greater overlap would demonstrate lower separability and specificity scores (negative correlation), along with higher homogeneity values (positive correlation), and vice versa. Our goal was to identify metrics with higher absolute mean correlation values and lower variability, as illustrated in Figures S21A and S21B. Based on simulation results (Figure S21C), Silhouette and bimodality index indicate high performance for the separability measure. The arithmetic mean of the average scaled variance (ASV) table shows higher performance compared to the geometric mean for the homogeneity measure. Our simulation results indicated a superior performance of simpson index compared to Shannon entropy[81] to measure factor specificity.

Calculating overlap between double Gaussian distributions:

The overlap between two Gaussian distributions with means μ_1 and μ_2, and standard deviations σ_1 and σ_2 respectively, is quantified based on the intersection of the PDFs of the two distributions. For distributions with unequal means and standard deviations, the overlap O is computed as:

$$O = 1 - \frac{1}{2}\left[\operatorname{erf}\left(\frac{c-\mu_1}{\sqrt{2}\sigma_1}\right) - \operatorname{erf}\left(\frac{c-\mu_2}{\sqrt{2}\sigma_2}\right)\right]$$

where c represents the intersection point of the two PDFs and erf is the error function.

Pathway and gene set enrichment analysis

Gene set and pathway enrichment analysis methods were used to study the biological signatures represented by each factor. The gene scores corresponding to the factors of interest were selected from the loading matrix to order the list of genes from most to least contribution to the given factor. Pathway enrichment analysis was performed on the top 200 genes of the ordered loadings using the gprofiler2[82] (version 0.2.3) enrichment tool based on the default Gene Ontology Biological Process and Reactome gene set database sources, and using the “ranked” mode.

Interpretation of factors using FCA and FIS tables

The FCA and FIS tables facilitate interpretation of identified factors. When covariates are available, evaluating the FCA table first can reveal factors associated with known covariates. Visual inspection of the FCA heatmap helps identify specific factor-covariate pairs, and the automatic thresholding highlights factor-covariate pairs with scores that stand out against the background distribution. Ideally, factors of interest will exhibit high specificity, as summarized in the FIS table. Specificity measures how uniquely a factor associates with a particular covariate level (e.g., scRNA-seq technology type), with higher specificity enhancing interpretability.

The FIS heatmap includes three additional metric scores - bimodality (separability), effect size, and homogeneity - which capture distinct aspects of the factor distribution and help prioritize unexplained factors for further exploration. Bimodality, which is covariate-independent, reflects the distribution’s

modality, where high values indicate the presence of two distinct cell populations with high and low factor scores. Effect size, also covariate-independent, represents the distribution's spread and variance. Higher effect size indicates higher signal strength. Homogeneity is covariate-dependent and measures the relative spread of the data specific to a single covariate compared to the total spread. Homogeneity values closer to one suggest cells are well-mixed with respect to the covariate along the factor's axis. These metrics guide the identification of interpretable factors from a set of K factors generated by any factorization algorithm. However, factor distributions may show patterns beyond these metrics, and even one high metric value can justify follow-up interpretation, guided by enriched genes and pathways or expert knowledge.

Modular design for flexible usage

sciRED's modular design allows integration with external data processing and normalization methods. Preprocessed data can be input directly into step 2 (factor identification). Additionally, if external matrix factorization has already been performed, the initial steps can be bypassed, with factors and covariates imported directly for factor-covariate association analysis.

Data preprocessing

scMixology

The scMixology[17] dataset includes three human lung adenocarcinoma cell lines: HCC827, H1975, and H2228. These cell lines were cultured individually and subsequently processed. Single cells from each cell line were combined in equal proportions and libraries were generated using three protocols: CEL-seq2, Drop-seq, and 10x Genomics Chromium. The processed count data was obtained using the scPipe package in R and converted to .h5ad for import into Python. The data underwent log normalization and standardization using the "StandardScaler" function from scikit-learn package.

Interferon (IFN)-β stimulated PBMC dataset

The stimulated data was downloaded from muscData package (Kang18_8vs8, GEO: GSE96583)[26] . This dataset includes 10x Genomics droplet-based scRNA-seq PBMC data from eight lupus patients before and after 6h-treatment with interferon-beta. Count data was extracted, and analyzed using sciRED (number of components(k)=30) while modeling library size as a technical covariate. Three outlier cells were removed from the sciRED cell-by-factor score matrix, and factors F2 and F9 scores for 29,062 cells were visualized in Figure 3.

Healthy human kidney atlas

The healthy human kidney map was constructed based on 19 living donors (10 female, 9 male)[28] including the transcriptomes of 27,677 cells. The processed files were downloaded from the UCSC Cell Browser [https://cells.ucsc.edu/?ds=living-donor-kidney]. Filtered and normalized data was downloaded and analyzed using sciRED (k=30).

Healthy rat liver atlas

The healthy rat total liver homogenate map includes four whole livers which were acquired from 8-10 week-old healthy male Dark Agouti and Lewis strain rats, and the resulting total liver homogenates went through two-step collagenase digestion and 10x Genomics droplet-based scRNA-seq[29]. The processed healthy rat liver total homogenate map was downloaded from the UCSC Cell Browser [https://cells.ucsc.edu/?ds=rat-liver-atlas] (GEO: GSE220075). Five outlier cells were removed from

the score matrix, and factors F6 and F20 scores for 23,036 cells were visualized in Figure 4. sciRED (k=30) was applied to the count data while modeling library size as a technical covariate. Sample was not included as technical covariate to preserve the strain-specific variations. SoupX[31] software (version: 1.6.2) was used to identify genes with the greatest contribution to the ambient RNA. We used the default automatic contamination fraction estimation (Rho) feature in the SoupX (autoEstCont function) to estimate Rho for each sample included in the total liver homogenate map of healthy rat livers. Subsequently, we extracted the estimated ambient RNA profile and identified the top 50 genes contributing the most to the ambient RNA in each sample. Genes were selected if they ranked among the high-scoring ambient RNA contributors in at least two samples. These selected genes were assessed for their presence among the strain-specific myeloid markers identified using both sciRED and standard differential expression methods. Differential expression analysis between the DA and LEW strains within the myeloid population (cluster 5) of the rat liver map was conducted using Seurat's FindMarkers function with default parameters (logfc.threshold = 0.1, min.pct = 0.01, min.cells.feature = 3, and min.cells.group = 3), implementing the non-parametric Wilcoxon rank-sum test.

Healthy human liver atlas

The healthy human liver map[32] includes 8,444 parenchymal and non-parenchymal cells obtained from the fractionation of fresh hepatic tissue from five human livers. The liver tissue was obtained from livers procured from deceased donors deemed acceptable for liver transplantation. Data was downloaded from the R package HumanLiver, available at https://github.com/BaderLab/HumanLiver, and sciRED (k=30) was applied to filtered count while modeling library size as a technical covariate.

Human lung transplant dataset

We downloaded the human lung transplants dataset[27] from the cellxgene platform (GEO: GSE220797), which includes donor lung biopsies from six transplant cases and over 108,000 cells. We then performed subsampling on both genes and cells to systematically assess sciRED's runtime across different dataset sizes.

PBMC 3' and 5' datasets

The two filtered 10k PBMC datasets, profiled using 10x Genomics single-cell 3' and 5' gene expression libraries, were downloaded directly from the 10x Genomics website (3p: https://cf.10xgenomics.com/samples/cell-exp/4.0.0/Parent_NGSC3_DI_PBMC/Parent_NGSC3_DI_PBMC_filtered_feature_bc_matrix.h5, 5p: https://cf.10xgenomics.com/samples/cell-vdj/5.0.0/sc5p_v2_hs_PBMC_10k/sc5p_v2_hs_PBMC_10k_filtered_feature_bc_matrix.h5).

Spatial transcriptomics data

We applied sciRED to spatial transcriptomics data from Maynard et al. (2020)[83], focusing on identifying spatial gene expression patterns within the six-layered human dorsolateral prefrontal cortex (DLPFC). Specifically, we selected Visium samples from two subjects (Br8100 and Br5292), with four samples per subject. The data was downloaded using spatialLIBD::fetch_data() and then subset by subject. sciRED's Poisson GLM was applied to adjust for library size.

Method availability

A Python package for sciRED with example case scenarios is freely available from: https://github.com/delipouya/sciRED

References

1. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3**, 1724–1735 (2007).
2. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol.* **21**, 31 (2020).
3. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* **95**, 14863–14868 (1998).
4. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).
5. Pearson, K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6* **2**, 559–572 (1901).
6. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).
7. Jutten, C. & Herault, J. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing* **24**, 1–10 (1991).
8. Comon, P. Independent component analysis, A new concept? *Signal Processing* **36**, 287–314 (1994).
9. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J.-P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* **9**, 284 (2018).
10. Pierson, E. & Yau, C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* **16**, 241 (2015).
11. Townes, F. W., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol.* **20**, 295 (2019).
12. Kotliar, D. *et al.* Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *eLife* **8**, (2019).
13. Stein-O'Brien, G. L. *et al.* Decomposing Cell Identity for Transfer Learning across Cellular Measurements, Platforms, Tissues, and Species. *Cell Syst.* **8**, 395-411.e8 (2019).
14. Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33**, 155–160 (2015).
15. Buettner, F., Pratanwanich, N., McCarthy, D. J., Marioni, J. C. & Stegle, O. f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biol.* **18**, 212 (2017).
16. Kunes, R. Z., Walle, T., Land, M., Nawy, T. & Pe'er, D. Supervised discovery of interpretable gene programs from single-cell data. *Nat. Biotechnol.* **42**, 1084–1095 (2024).
17. Tian, L. *et al.* Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat. Methods* **16**, 479–487 (2019).
18. Nelder, J. A. & Wedderburn, R. W. M. Generalized Linear Models. *J. R. Stat. Soc. Ser. A* **135**, 370 (1972).
19. Kaiser, H. F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **23**, 187–200 (1958).

20. Cox, D. R. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)* **20**, 215–232 (1958).
21. Bamber, D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J. Math. Psychol.* **12**, 387–415 (1975).
22. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **27**, 861–874 (2006).
23. Wu, X. *et al.* Top 10 algorithms in data mining. *Knowl. Inf. Syst.* **14**, 1–37 (2008).
24. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* 785–794 (ACM Press, 2016). doi:10.1145/2939672.2939785.
25. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
26. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
27. Wong, A. *et al.* Ischemia-reperfusion responses in human lung transplants at the single-cell resolution. *Am. J. Transplant.* (2024) doi:10.1016/j.ajt.2024.08.019.
28. McEvoy, C. M. *et al.* Single-cell profiling of healthy human kidney reveals features of sex-based transcriptional programs and tissue-specific immunity. *Nat. Commun.* **13**, 7634 (2022).
29. Pouyabahar, D. *et al.* A rat liver cell atlas reveals intrahepatic myeloid heterogeneity. *iScience* **26**, 108213 (2023).
30. Andrews, T. S. *et al.* Single cell, single nucleus and spatial RNA sequencing of the human liver identifies hepatic stellate cell and cholangiocyte heterogeneity. *BioRxiv* (2021) doi:10.1101/2021.03.27.436882.
31. Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* **9**, giaa151. (2020).
32. MacParland, S. A. *et al.* Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat. Commun.* **9**, 4383 (2018).
33. Andrews, T. S. *et al.* Single-cell, single-nucleus, and spatial transcriptomics characterization of the immunological landscape in the healthy and PSC human liver. *J. Hepatol.* **80**, 730–743 (2024).
34. Halpern, K. B. *et al.* Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature* **542**, 352–356 (2017).
35. Nicolau-Neto, P. *et al.* UBE2C is a transcriptional target of the cell cycle regulator FOXM1. *Genes* **9**, (2018).
36. Wang, T., Lu, J., Wang, R., Cao, W. & Xu, J. TOP2A promotes proliferation and metastasis of hepatocellular carcinoma regulated by miR-144-3p. *J. Cancer* **13**, 589–601 (2022).
37. Huang, L. *et al.* KPNA2 promotes cell proliferation and tumorigenicity in epithelial ovarian carcinoma through upregulation of c-Myc and downregulation of FOXO3a. *Cell Death Dis.* **4**, e745 (2013).
38. Martinsson-Ahlzén, H.-S. *et al.* Cyclin-dependent kinase-associated proteins Cks1 and Cks2 are essential during early embryogenesis and for cell cycle progression in somatic cells. *Mol. Cell. Biol.* **28**, 5698–5709 (2008).
39. Xu, L., Yu, W., Xiao, H. & Lin, K. BIRC5 is a prognostic biomarker associated with tumor immune cell infiltration. *Sci. Rep.* **11**, 390 (2021).
40. Cao, Y., Yang, P. & Yang, J. Y. H. A benchmark study of simulation methods for single-cell RNA sequencing data. *Nat. Commun.* **12**, 6911 (2021).
41. Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data.

Genome Biol. **18**, 174 (2017).
42. Li, W. V. & Li, J. J. A statistical simulator scDesign for rational scRNA-seq experimental design. *Bioinformatics* **35**, i41–i50 (2019).
43. Vieth, B., Ziegenhain, C., Parekh, S., Enard, W. & Hellmann, I. powsimR: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics* **33**, 3486–3488 (2017).
44. Baruzzo, G., Patuzzi, I. & Di Camillo, B. SPARSim single cell: a count data simulator for scRNA-seq data. *Bioinformatics* **36**, 1468–1475 (2020).
45. Zhang, X., Xu, C. & Yosef, N. Simulating multiple faceted variability in single cell RNA sequencing. *Nat. Commun.* **10**, 2611 (2019).
46. Stein-O'Brien, G. L. *et al.* Enter the Matrix: Factorization Uncovers Knowledge from Omics. *Trends Genet.* **34**, 790–805 (2018).
47. Revelle, W. psych: Procedures for Psychological, Psychometric, and Personality Research. (2024).
48. Waller, N. simFA: A Program for Exploring Factor Analysis Models Through Monte Carlo Simulations. (2021).
49. Briggs, N. E. & MacCallum, R. C. Recovery of weak common factors by maximum likelihood and ordinary least squares estimation. *Multivariate Behav. Res.* **38**, 25–56 (2003).
50. Cliff, N. & Pennell, R. The influence of communality, factor strength, and loading size on the sampling characteristics of factor loadings. *Psychometrika* **32**, 309–326 (1967).
51. Svensson, V., Gayoso, A., Yosef, N. & Pachter, L. Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics* **36**, 3418–3421 (2020).
52. Choi, Y., Li, R. & Quon, G. siVAE: interpretable deep generative models for single-cell transcriptomes. *Genome Biol.* **24**, 29 (2023).
53. Jiang, R., Sun, T., Song, D. & Li, J. J. Statistics or biology: the zero-inflation controversy about scRNA-seq data. *Genome Biol.* **23**, 31 (2022).
54. Gastwirth, J. L. The estimation of the lorenz curve and gini index. *Rev. Econ. Stat.* **54**, 306 (1972).
55. Lause, J., Berens, P. & Kobak, D. Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. *Genome Biol.* **22**, 258 (2021).
56. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).
57. Pan, Y. *et al.* The Poisson distribution model fits UMI-based single-cell RNA-sequencing data. *BMC Bioinformatics* **24**, 256 (2023).
58. Seabold, S. & Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. in *Proceedings of the 9th Python in Science Conference* 92–96 (SciPy, 2010). doi:10.25080/Majora-92bf1922-011.
59. Klema, V. & Laub, A. The singular value decomposition: Its computation and some applications. *IEEE Trans. Automat. Contr.* **25**, 164–176 (1980).
60. Varoquaux, G. *et al.* Scikit-learn. *GetMobile: Mobile Comp. and Comm.* **19**, 29–33 (2015).
61. Stegmann, M. B., Sjöstrand, K. & Larsen, R. Sparse modeling of landmark and texture variability using the orthomax criterion. in *Medical Imaging 2006: Image Processing* (eds. Reinhardt, J. M. & Pluim, J. P. W.) vol. 6144 61441G (SPIE, 2006).
62. Dien, J., Khoe, W. & Mangun, G. R. Evaluation of PCA and ICA of simulated ERPs: Promax vs. Infomax rotations. *Hum. Brain Mapp.* **28**, 742–763 (2007).
63. Hendrickson, A. E. & White, P. O. PROMAX: A QUICK METHOD FOR ROTATION TO OBLIQUE SIMPLE STRUCTURE. *British Journal of Statistical Psychology* **17**, 65–70 (1964).

64. Akhtar-Danesh, N. Impact of factor rotation on Q-methodology analysis. *PLoS ONE* **18**, e0290728 (2023).
65. Darton, R. A. Rotation in factor analysis. *The Statistician* **29**, 167 (1980).
66. Tin Kam Ho. Random decision forests. in *Proceedings of 3rd International Conference on Document Analysis and Recognition* 278–282 (IEEE Comput. Soc. Press, 1995). doi:10.1109/ICDAR.1995.598994.
67. Cover, T. & Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory* **13**, 21–27 (1967).
68. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
69. Mitchell, D. W. 88.27 More on spreads and non-arithmetic means. *The Mathematical Gazette* **88**, 142–144 (2004).
70. Otsu, N. A Threshold Selection Method from Gray-Level Histograms. *IEEE Trans. Syst. Man Cybern.* **9**, 62–66 (1979).
71. van der Walt, S. *et al.* scikit-image: image processing in Python. *PeerJ* **2**, e453 (2014).
72. Hellwig, B. *et al.* Comparison of scores for bimodality of gene expression distributions and genome-wide evaluation of the prognostic relevance of high-scoring genes. *BMC Bioinformatics* **11**, 276 (2010).
73. Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inform. Theory* **28**, 129–137 (1982).
74. Calinski, T. & Harabasz, J. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods* **3**, 1–27 (1974).
75. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 53–65 (1987).
76. Davies, D. L. & Bouldin, D. W. A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1**, 224–227 (1979).
77. Wang, J., Wen, S., Symmans, W. F., Pusztai, L. & Coombes, K. R. The bimodality index: a criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data. *Cancer Inform.* **7**, 199–216 (2009).
78. Hartigan, J. A. & Hartigan, P. M. The Dip Test of Unimodality. *Ann. Statist.* **13**, 70–84 (1985).
79. Simpson, E. H. Measurement of Diversity. *Nature* **163**, 688–688 (1949).
80. Hsu, L. L. & Culhane, A. C. Correspondence analysis for dimension reduction, batch integration, and visualization of single-cell RNA-seq data. *Sci. Rep.* **13**, 1197 (2023).
81. Shannon, C. E. A mathematical theory of communication. *Bell System Technical Journal* **27**, 379–423 (1948).
82. Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* **35**, W193-200 (2007).
83. Maynard, K. R. *et al.* Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat. Neurosci.* **24**, 425–436 (2021).

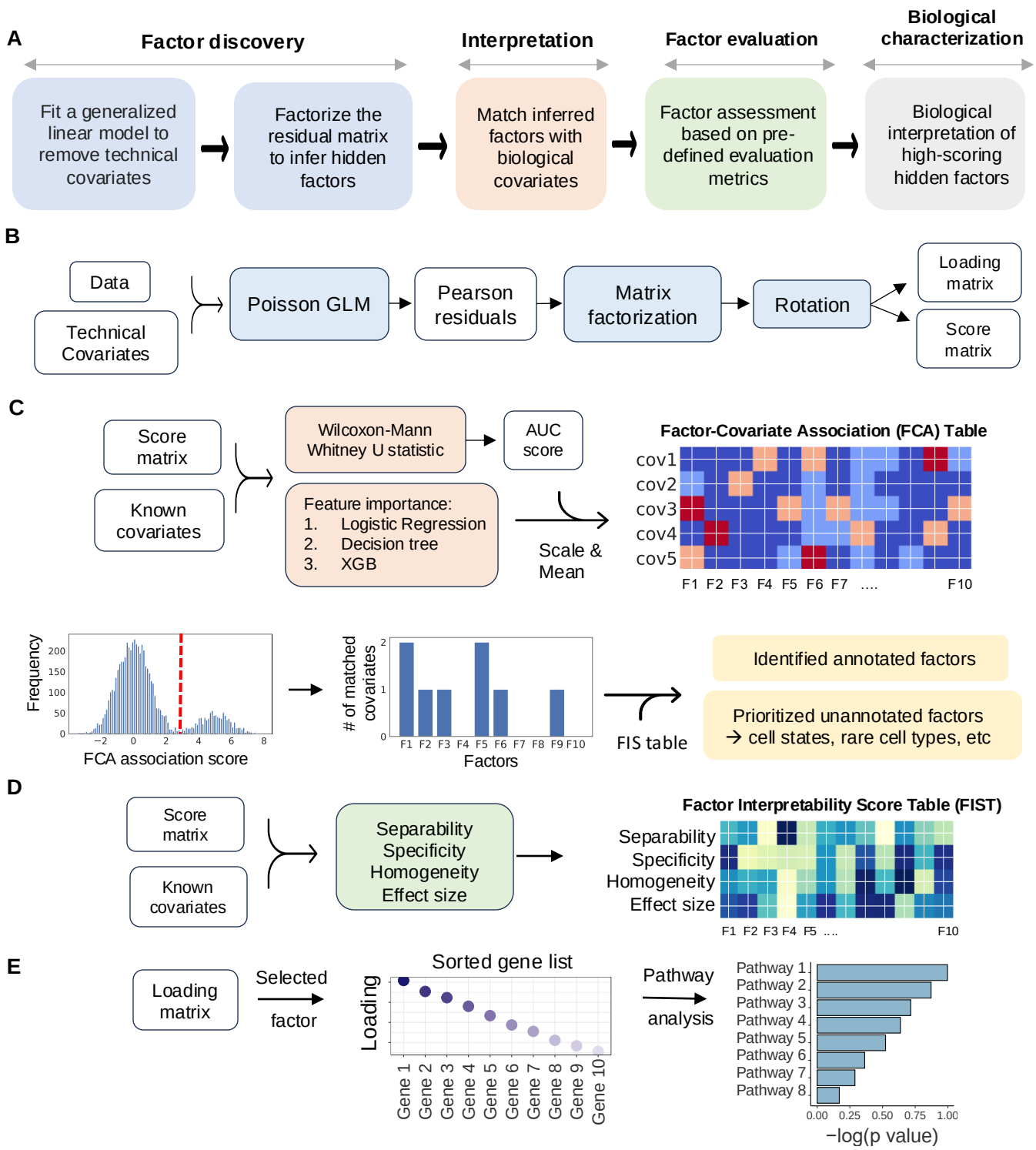

Figure 1) sciRED overview.

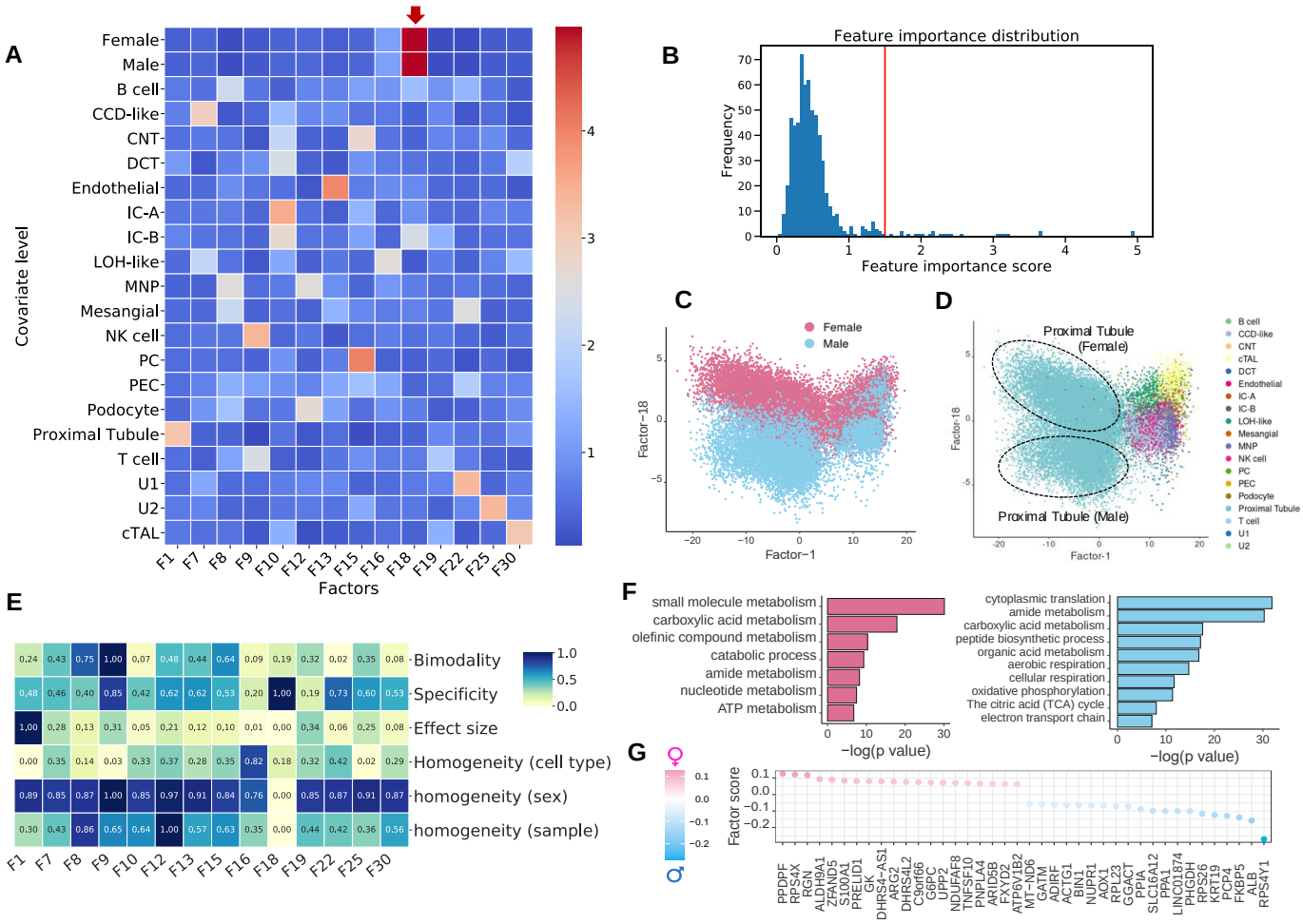

Figure 2) Deciphering cell type identity programs and sex-specific processes in a human healthy kidney map.

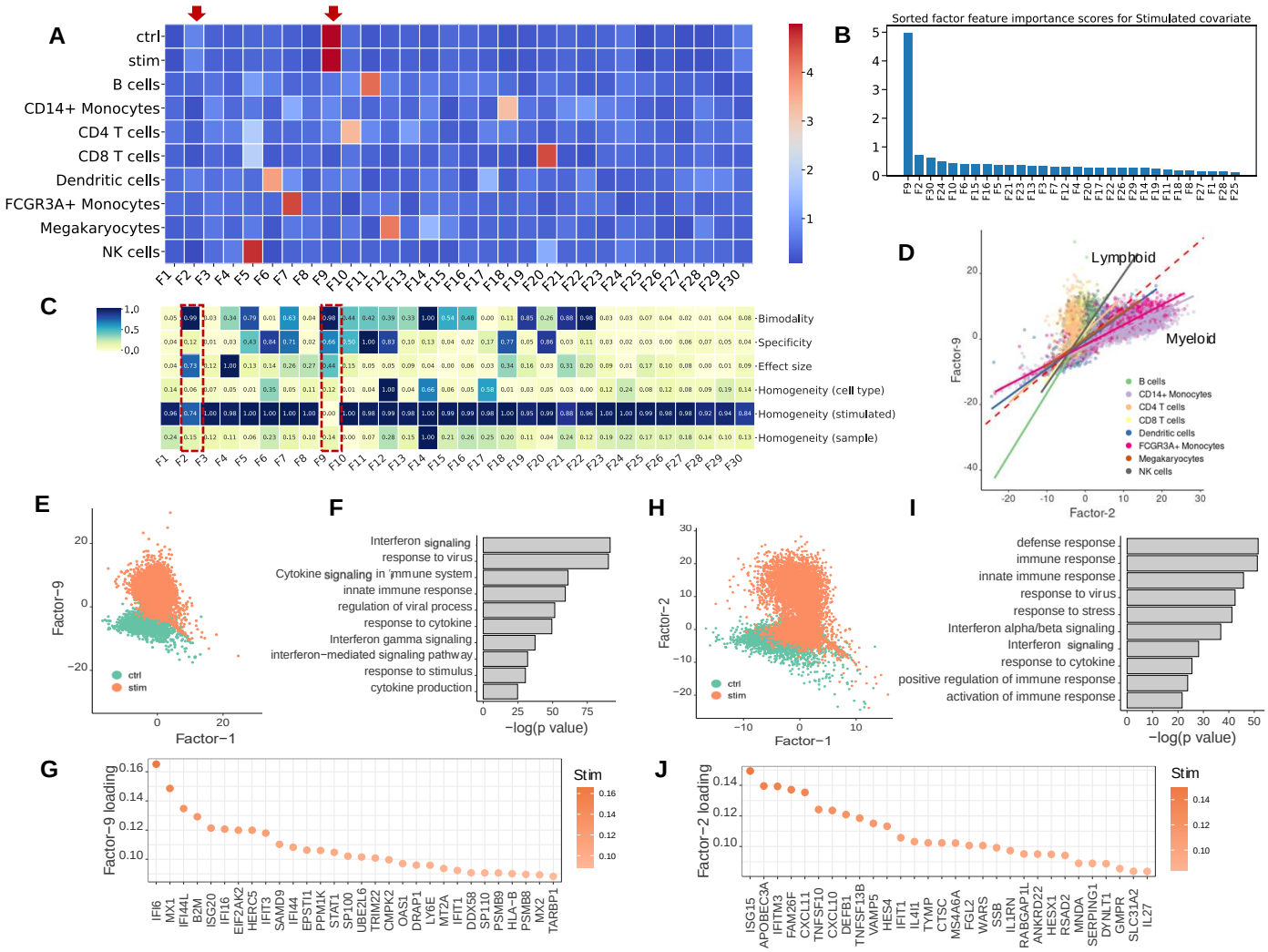

Figure 3) Identifying interferon-β stimulation signals across lymphoid and myeloid cells in a PBMC dataset.

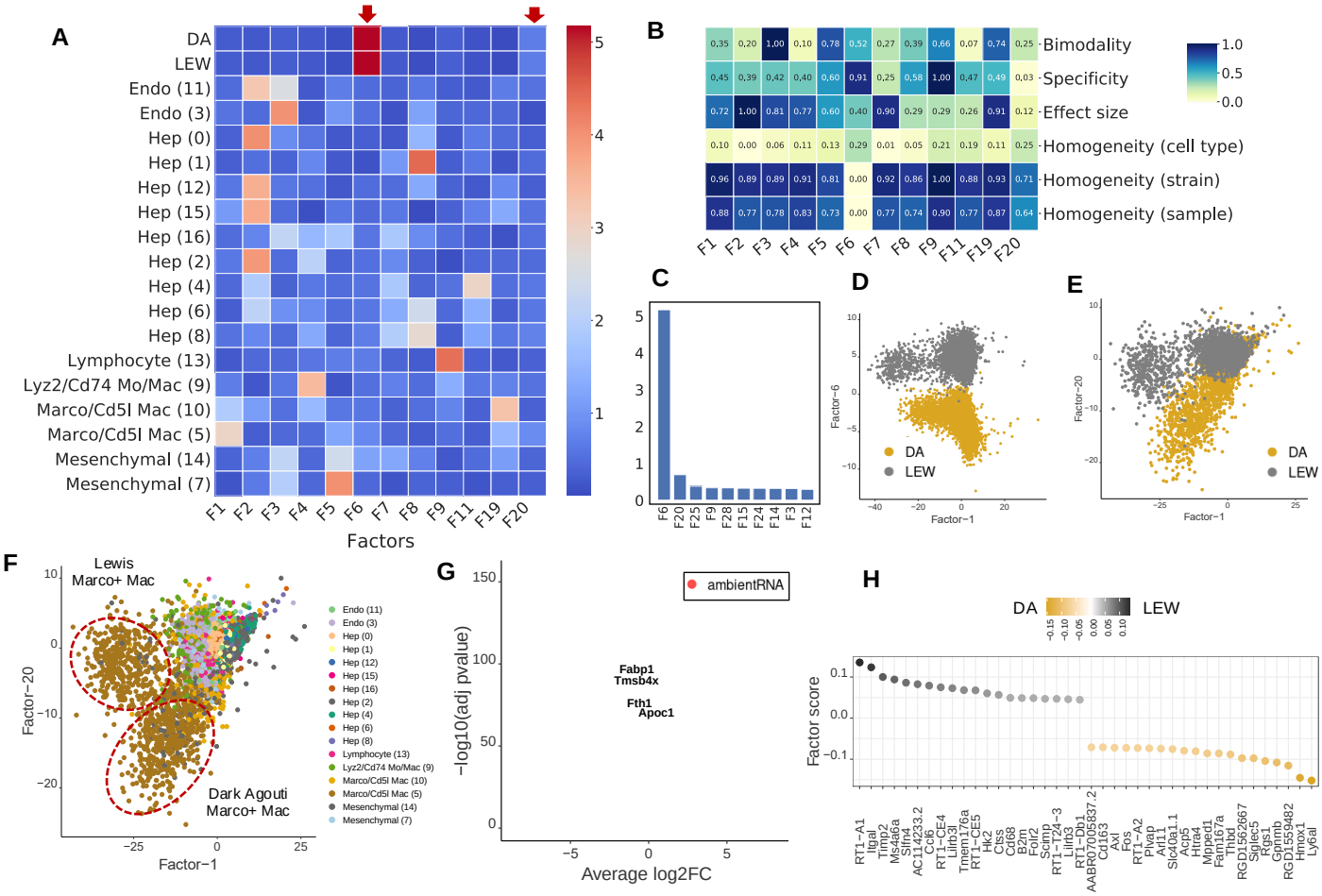

Figure 4) sciRED identifies strain-based variation despite ambient RNA contamination in a rat liver map.

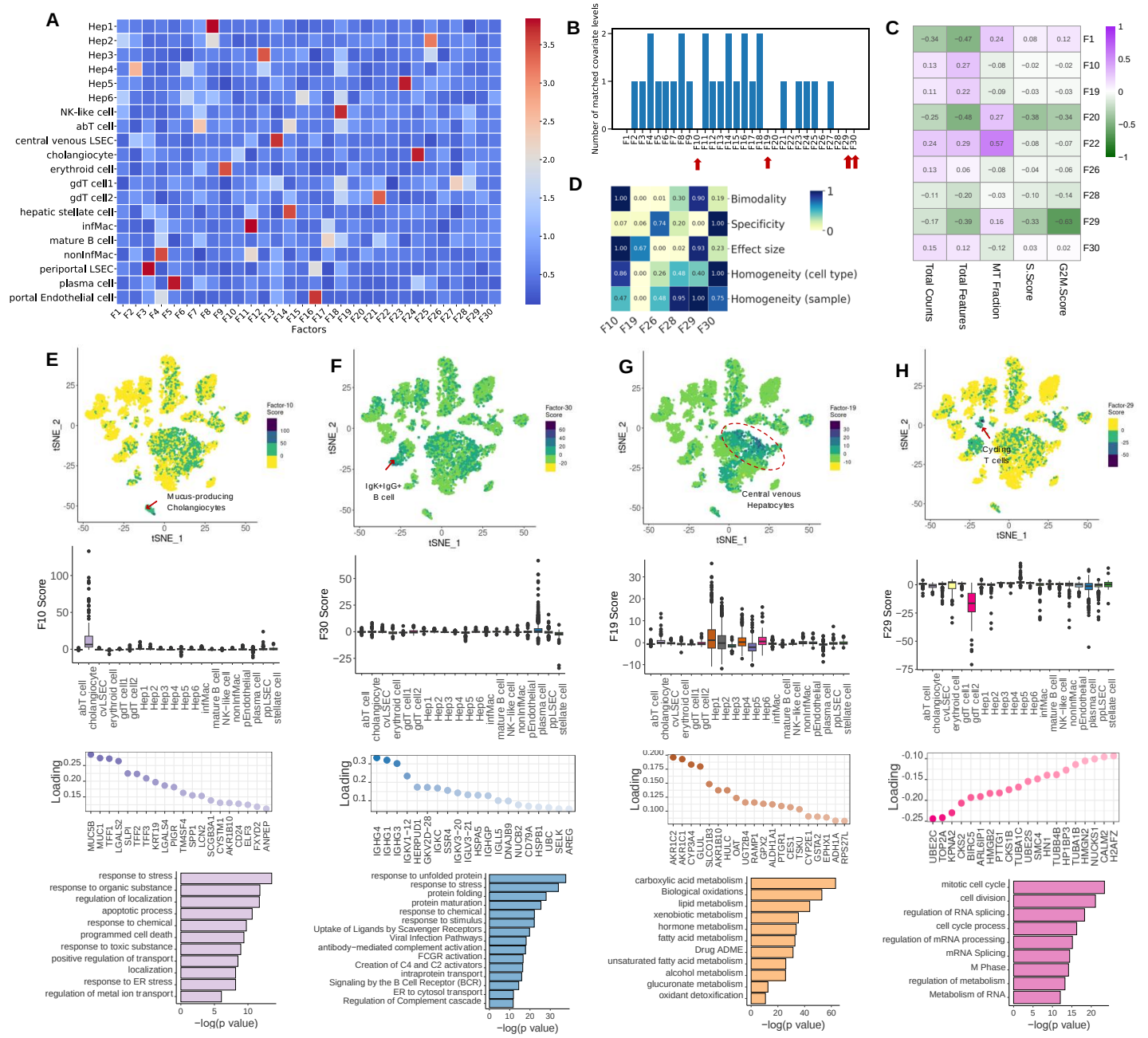

Figure 5) Exploring hidden biology in the healthy human liver atlas using unannotated factors.