Joint Modeling of Cellular Heterogeneity and Condition Effects with scPCA in Single-Cell RNA-Seq

Harald Vöhringer[1-4] and Sascha Dietrich[2,3]

[1] European Molecular Biology Laboratory (EMBL), Heidelberg, Germany
[2] Molecular Medicine Partnership Unit (MMPU), Heidelberg, Germany
[3] Department of Medicine V, Hematology, Oncology and Rheumatology, University Hospital Heidelberg, Heidelberg, Germany
[4] Department of Hematology and Oncology, University Hospital Düsseldorf, Düsseldorf, Germany

Abstract

Single-cell RNA sequencing (scRNA-seq) in multi-condition experiments enables the systematic assessment of treatment effects. Analyzing scRNA-seq data relies on linear dimensionality reduction (DR) methods like principal component analysis (PCA). These methods decompose high-dimensional gene expression profiles into tractable factor representations and prototypical gene expression patterns (components), facilitating the study of cell type variation. However, integrating study covariates within linear DR frameworks remains a challenging task. We present scPCA, a flexible DR framework that jointly models cellular heterogeneity and conditioning variables, allowing it to recover an integrated factor representation and reveal transcriptional changes across conditions and components of the decomposition. We show that scPCA extracts an interpretable latent representation by analyzing unstimulated and IFN-beta-treated PBMCs, and showcase that the model may be employed to effectively address batch effects. We examine age-related changes in rodent lung cell populations, uncovering a previously unreported surge in *Ccl5* expression in T cells. We illustrate how scPCA may be employed to identify coordinated transcriptional changes across multiple time-points in depolarized visual cortex neurons. Finally, we show that scPCA elucidates transcriptional shifts in CRISPR-Cas9 chordin knockout zebrafish fish single-cell data despite large difference cell abundance across conditions. Since scPCA introduces a general approach to account for conditioning variables in high-dimensional data, it may also be applicable to datasets other than scRNA-seq.

Correspondence

Harald Vöhringer, PhD
Universitätsklinik Heidelberg and EMBL
Meyhofer Str. 1
69117 Heidelberg
harald.voehringer@embl.de

Introduction

Multi-condition single-cell RNA sequencing (scRNA-seq) provides a comprehensive way to analyze how treatments impact various cell types systematically. However, integrating and interpreting single-cell data is challenging, especially if the biological question requires assessing cells across multiple conditions and technical replicates. This complexity stems from the inherently high-dimensional nature of the data and the influence of two primary sources of variation: cellular heterogeneity and experimental perturbation[1].

Central to the analysis workflow for scRNA-seq data are linear dimensionality reduction (DR) methods. These methods decompose high-dimensional gene expression matrices into two smaller matrices: the factor matrix, which provides a low-dimensional representation of cells used for clustering[2] and data visualization techniques[3,4], and the loading matrix, which contains basis vectors or components capturing prototypical expression patterns. However, DR methods like PCA struggle to disentangle variation from cellular heterogeneity and experimental intervention[5,6], often requiring post-hoc adjustments to the factors[6,7]. While some linear DR methods model conditioning variables, they were not intended to capture fully integrated factors, meaning they fail to represent the same cell types from different conditions with similar latent representations. Additionally, these methods do not generalize to experiments with multiple conditioning variables[8] (e.g. genetic perturbation and drug stimulation).

For example, contrastive PCA[9,10] identifies components by contrasting the variance in a target dataset against a background dataset, thus highlighting unique patterns specific to the target. Multi-Omics Factor Analysis (MOFA+)[11] uses sparsity priors to identify subsets of factors that explain condition-specific observations. Both approaches identify condition-specific factors but do not examine how components vary across conditions. The Gene Expression Decomposition and Integration (GEDI)[8] framework enables the extraction of an integrated factor representation by accounting for component changes across conditions. However, GEDI's fixed parameterisation, which may be considered as a specific case of the more general method presented here, may only account for a single conditioning variable at a time (for a more comprehensive overview see also the **Supplementary Methods**).

The problem of disentangling latent spaces with respect to conditioning variables has been explored in the context of non-linear low-rank representation learning[12]. For example, conditional variational autoencoders[13] (cVAEs) retrieve integrated data representations by informing the model about the conditioning variable[14–16]. However, unlike linear DR methods, where each component can be interpreted as a "metagene", the encoder and decoder functions in cVAEs are less straightforward to interpret.

We developed scPCA, a flexible factorization model for analyzing multi-condition single-cell datasets. Our model incorporates conditioning variables, enabling scPCA to extract condition-specific bases. This allows analysts to assess expression shifts across each condition and component and retrieves an integrated factor representation of the data. We exemplify how scPCA extracts a low-rank representation that captures the main drivers of variation by applying our tool to untreated and IFN beta-stimulated PBMCs[17]. We showcase

how our factorisation model may be employed to address batch effects by integrating cell-line data, and assess it in various single-cell batch integration benchmarking tasks[18]. We illustrate that scPCA factors often discriminate cell-type clusters by analyzing lung cells of aging mice, and demonstrate that their corresponding components may be leveraged to study their constituent gene programs. We then apply scPCA to a dataset comprising multiple timepoints to reveal early and late transcriptional responses in various cell populations of depolarized rodent brain cells. Finally, we show that scPCA faithfully recovers the effects of a CRISPR-Cas9 chordin knockout in zebrafish embryos despite uneven cell type distributions across conditions.

Results

Disentangling experimental states and cellular heterogeneity

In scPCA, we consider single-cell expression count matrices with categorical covariates indicating conditions such as drug treatment[17] or genetic perturbation[19,20] (**Fig. 1a, b**). These conditioning variables are represented as a full-rank design matrix, used to parameterize a loading weight tensor that captures the condition-specific principal axes of variation relative to a user-defined reference condition (**Fig 1c, Methods**). Additionally, we model a mean offset matrix, which can optionally be parameterized using an indicator matrix, allowing scPCA to account for condition-specific shifts in the data space. This adjustment is sometimes necessary because different conditions can cause global shifts in gene expression profiles, leading to distinct data distributions that must be properly aligned to accurately extract the principal components governing the data (**Fig. 1c**, **Methods**).

Multi-condition single-cell datasets are characterized by shared cell types that exhibit altered expression profiles as a result of experimental interventions. For this reason, we hypothesized that such data are governed by closely related but not identical lower-dimensional principal axes of variation. To illustrate this concept, we generated multi-condition data in $\mathbb{R}^2$ using two distinct 1-dimensional bases to model condition-specific data subspaces, and simulated two clusters to represent two shared cell types (**Fig. 1d**, **Methods**). Note that cell cluster A from the treatment (trt) condition has a higher propensity to express gene x_2 in comparison to its counterpart from the reference group (**Fig. 1d**). This demonstrates how systematic shifts in gene expression across one or multiple genes can introduce additional variability within cell type clusters, even when the same cell types are present across different conditions (**Fig. 1d**).

Modeling single-cell data without explicitly accounting for its experimental structure leads to the extraction of a basis that fails to reconstruct the true data-generating subspaces (**Fig. 1e**). This results in large residuals between observed and reconstructed data points (**Fig. 1e**, gray lines). Consequently, DR methods explain the variation induced by the condition with additional factors leading to a lower-dimensional representation of the data that conflates variation due to cellular heterogeneity with variability attributed to the experimental intervention. Notably, the existence of such factors impede subsequent downstream analyses, given that the factor representation serves as the foundation for downstream tasks such as clustering[21,22] and visualization[3,4].

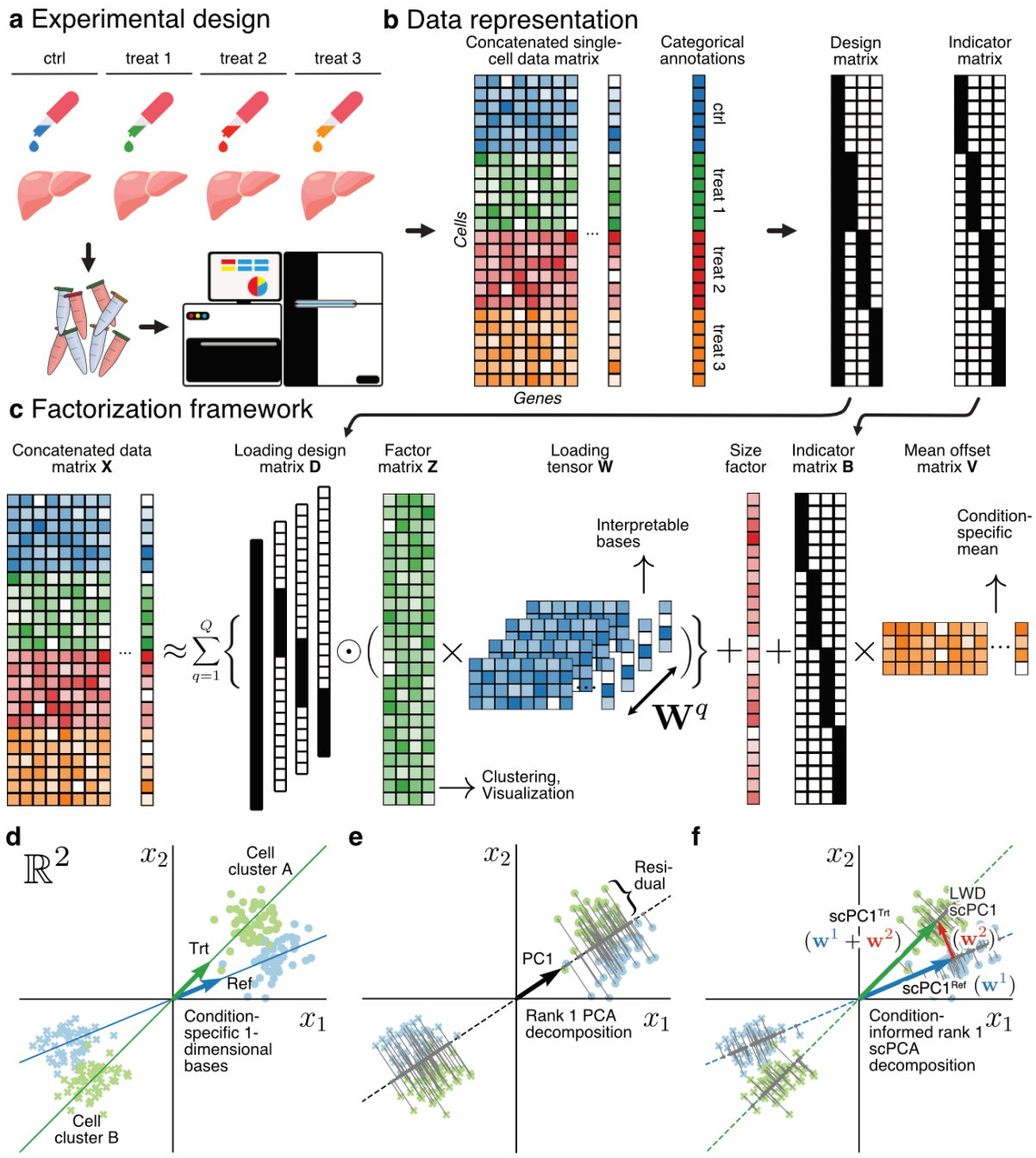


Fig 1: scPCA's factorisation framework allows to account transcriptional changes in expression across experimental conditions. a. Schematic depiction of an experimental design involving multiple treatments. **b.** A single-cell expression count matrix may be obtained by concatenating the expression matrices of each experimental condition. Categorical covariates may be leveraged in scPCA in the form of design and indicator matrices. **c.** Schematic depiction of the factor model (the $\odot$ symbol denotes element-wise multiplication). The loading design matrix parameterizes a loading tensor, enabling the model to identify condition-specific bases. An optional indicator matrix may be employed to fit different mean offsets to each experimental condition. **d.** Schematic depiction of simulated single-cell data in 2-dimensional (log) gene space. Arrows and lines indicate the data-generating 1-dimensional bases and subspaces of the reference (blue) and treatment (green) condition, respectively. **e.** Ignoring the conditioning variable leads to the extraction of

a basis that fails to reconstruct the principal subspaces of the reference or treated condition. As a consequence, conventional factor models such as PCA explain the residual variance with additional factors (not shown). **f.** Incorporating the loading design matrix in scPCAs factorisation framework enables the extraction of condition-specific basis vectors (scPCs). In **c**, **d** and **e** dots and crosses indicate two cell clusters. Green and blue data points highlight the reference and treatment condition.

Conversely, in scPCA, we factor out the variability attributed to the experimental intervention by allowing the model to extract distinct, yet aligned condition-specific basis vectors (hereafter scPCs). This results in a more accurate representation of the data-generating subspaces, and thus to smaller residuals between observed data points and corresponding projections (**Fig. 1f**). As a result, scPCA factors explain a greater proportion of the variance compared to conventional factor models at equivalent decomposition ranks and eliminate the necessity for factors that explicitly address variance attributed to the conditioning variable (**Fig. 1f**). Finally, due to the linearity of the model, a scPCA decomposition enables the analyst to assess the change of individual scPCs across conditions, thereby providing insights into how the components vary under different conditions.

We implemented scPCA as a hierarchical Bayesian factor model and employ stochastic variational inference to guarantee scalability to large single-cell datasets[23–25]. To ensure robust inference, we employ an overdispersed Negative Binomial distribution to model the count expression matrix (**Methods**)[26,27], and tested the model extensively in simulation experiments (**Supplementary Fig. 1d-f**). Taken together, scPCA's flexible factorisation framework effectively deconvolves the principal axes of variation with respect to experimental covariates.

Accounting for interferon-beta stimulation in PBMCs

To assess the efficacy of scPCA in integrating and interpreting multi-condition single-cell datasets, we employed our tool on public scRNA-seq data of PBMCs obtained from 8 lupus patients. These cells were either activated with recombinant IFN beta (stim) or left untreated (ctrl) for a duration of six hours. The dataset comprises a total of 24,305 cells, roughly divided into equal parts of treated (12,138) and untreated (12,167) single cells[17].

The strong perturbation effect of IFN beta induces widespread transcriptomic alterations in all cells[5]. As a result, applying the standard scRNA-seq analysis pipeline, which included PCA and uniform manifold approximation and projection[3] (UMAP), led to cell clustering based not only on cell type but also on condition, complicating the interpretation of the data (**Fig. 2a**). We then analyzed the PCA factors, which revealed that the first factor determined the lymphoid-myeloid axis. However, the second factor separated the cells by condition, explaining the poor integration of the data (**Fig. 2d**).

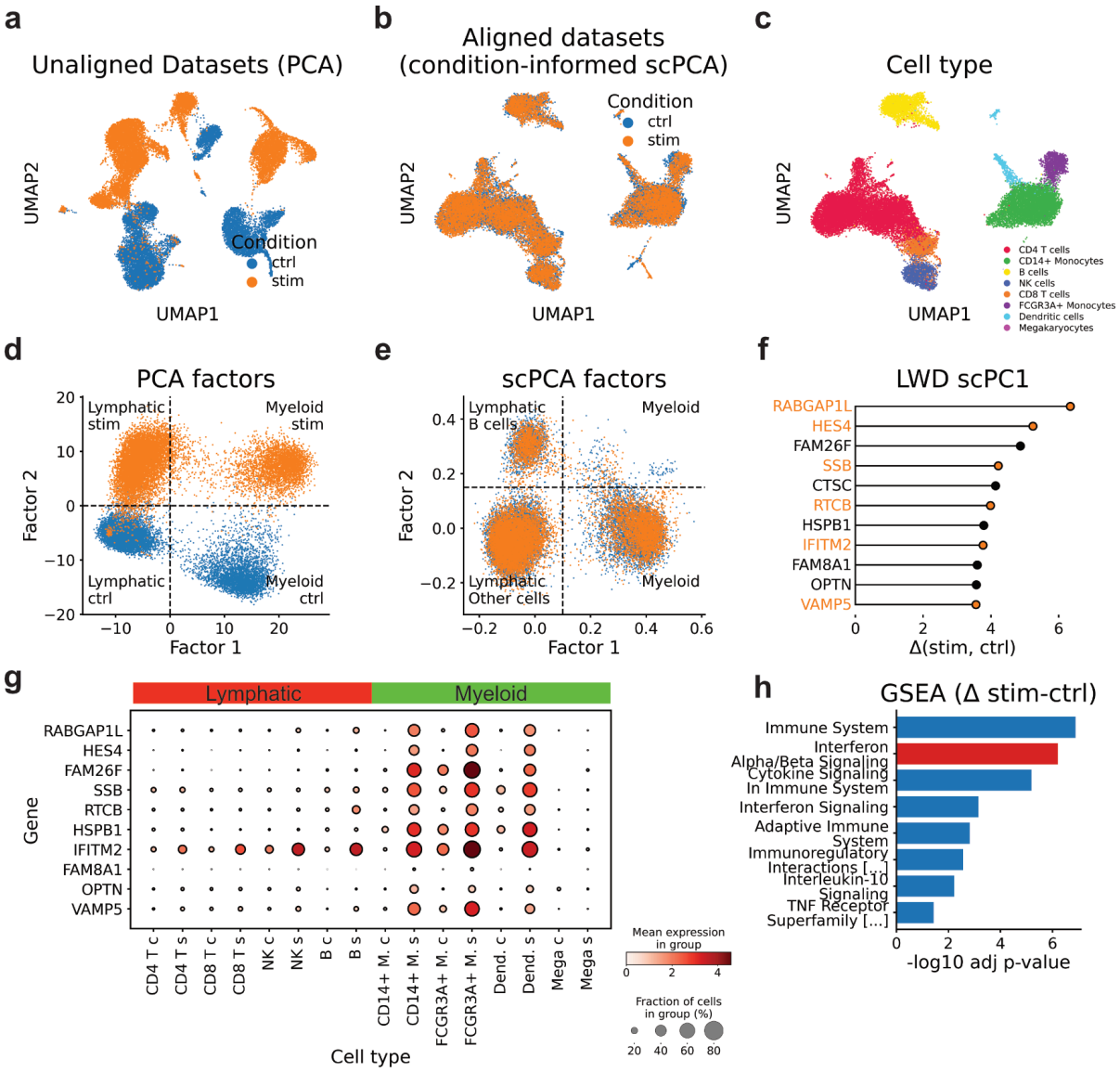


Fig 2: Modeling control and IFN beta stimulated PBMCs with scPCA. a. UMAP visualization showing untreated (control) and IFN-beta stimulated (stim) PBMCs (n=24,305), derived from a rank 20 PCA decomposition. **b. and c.** UMAP plots generated from a condition-informed rank 20 scPCA decomposition, with annotations for condition (**b**) and cell types (**c**). **d. and e.** Scatter plots illustrating the first two factors from PCA (**d**) and scPCA (**e**). **f.** Positive scPC1 LWDs indicating genes with the propensity to be higher expressed in stimulated myeloid cells. **g.** Dot plot showing genes with largest scPC1 LWDs (**f**) across cell types and conditions (the abbreviations c and s denote control and stimulated cells, respectively). **h.** A gene set enrichment analysis (GSEA) based on the top LWDs of scPC1 indicates upregulated gene programs in IFN beta-stimulated myeloid cells, including IFN-beta signaling (highlighted).

Subsequently, we applied a condition-informed scPCA decomposition to the data, allowing the model to fit condition-specific components (**Supplementary Fig. 2a, Methods**). Notably, the UMAP based on scPCA factors indicated a successful integration of the dataset (**Fig. 2b**, **c**), which was further supported by integration and biological conservation scores

(**Supplementary Fig. 2h**)[6,18]. Akin to PCA, we observed that the first scPCA factor separated cells based on lineage (**Fig. 2e**). However, the second scPCA factor explained variance within lymphatic cells rather than accounting for variation associated with the condition (**Fig. 2e**), indicating that scPCA successfully accounted for the variation induced by the drug stimulation.

Next, we investigated whether our algorithm produced a latent representation comparable to PCA when information about the condition is omitted (**Supplementary Fig. 2b, Methods**). Similarly to PCA, we found that scPCA's cell embeddings generated an UMAP representation wherein ctrl and stim cells separated by condition (**Supplementary Fig. 2c**). Furthermore, we verified that the first two factors explained lineage and IFN-stimulation (**Supplementary Fig 2d**), establishing these variables as major drivers of variation in the dataset.

We then analyzed scPCA's loading tensor to better understand how the model accounted for the two conditions. The cross-correlation of components revealed minimal similarity among the scPCs, implying that the basis vectors were largely orthogonal to each other (**Supplementary Fig 2e**, off-diagonal elements). However, a closer inspection of the same component across conditions (diagonal terms) revealed high correlations, suggesting that the scPCs were largely preserved across conditions (**Supplementary Fig. 2f**, diagonal elements). These observations indicate that while different scPCs capture largely orthogonal sources of variation, the same scPC captures similar sources of variation across conditions.

We proceeded to analyze the first components of the scPCA factorization. Subjecting the top gene weights of scPC1 to gene set enrichment analysis (GSEA) revealed associations with gene sets related to the innate immune system (**Supplementary Fig. 2g**), aligning with the enrichment of corresponding factor weights in myeloid cells (**Fig. 2e**). Next, to understand how IFN beta altered the gene expression profile of myeloid cells, we computed scPC1 loading weight differences (LWDs) across conditions (**Fig. 2f**, **Methods**). This analysis revealed the expected upregulation of genes involved in cytokine-mediated pathways, including *IFITM2*, *ISG20* and *CXCL10*. Notably, largest LWDs were observed in genes such as *RABGAP1L*, *HES4*, and *FAM26F*, which are associated with protein translocation (**Fig. 2f**). This suggests altered protein metabolism in myeloid cells as a result of the stimulation[28]. The analysis of gene expression confirmed that the mentioned genes were upregulated in stimulated myeloid cells, but not in other cell types (**Fig. 2g**). Finally, a GSEA based on scPC1 LWDs revealed an activated IFN-beta process, providing evidence that LWDs accurately capture activated gene programs (**Fig. 2h**). Taken together, these observations suggest that LWDs capture cell-type-specific changes in gene expression across conditions.

The flexible scPCA framework enables to model batch effects

Single-cell RNA-seq studies often comprise multiple experiments with differences in capturing times, experimenters, reagents, equipment and technology platforms[29]. These variations can introduce batch effects, reflecting variability due to technical artifacts rather than biological differences. There is a large body of literature on computational methods addressing the challenge of integrating such datasets, which rely on approaches ranging from mutual nearest neighbors[1], matrix factorisation[6,30], or more complex models based on

neural networks[14,31]. However, none of the aforementioned methods offer an interpretable explanation for batch effects, leading to concerns about potential overcorrection[18,29].

We hypothesized that scPCA's ability to incorporate conditioning variables could help to account for batch variation in single-cell data. For this reason, we used previously published single-cell data[32], comprising a 50/50 mixture (n=3,364) of Jurkat and t293 cells, and pure batches of Jurkat (n=5,054) and t293 (n=2,859) cells, and attempted to model them jointly using scPCA. Notably, Jurkat cells, which are derived from a male cell line, predominantly express the *CD3D* gene, whereas t293 cells, originating from a female cell line, exhibit a pronounced expression of the X-chromosome silencing gene *XIST*[32].

As previously noted[6], using PCA embeddings as input for UMAP resulted in unsatisfactory outcomes, wherein only t293 cells from the mixture and pure batch integrate (**Fig. 3a**). A closer examination of PCA factors indicated that the first factor was likely to distinguish both cell lines, while the second factor had the tendency to assign positive and negative weights to Jurkat cells from the pure and half batch, respectively (**Fig. 3d**).

Next, we applied scPCA to the dataset by selecting the half-batch as reference condition (**Supplementary Fig. 3b**). Our findings indicated that the scPCA factorization effectively addressed the batch effect, as Jurkat and t293 cells from the mixed and the pure batches mixed in the UMAP representation of the data (**Fig. 3b, c**) and exhibited better integration metrics in comparison to PCA (**Supplementary Fig. 3e**). Further, we confirmed the lack of a factor with a clear tendency to explain the batch structure (**Fig. 3e**), thus providing an explanation for the successful data integration.

We then ran scPCA in a batch-agnostic manner to assert that the model extracts a latent representation akin to PCA when information about the batch structure is omitted (**Supplementary Fig. 3b**, **Methods**). We found that the UMAP constructed from the uninformed scPCA model failed to account for the batch effect (**Supplementary Fig. 3c**). Similar to PCA, the analysis of factors indicated that the second largest source of variation was likely due to differences across batches in Jurkat cells (**Supplementary Fig. 3d**).

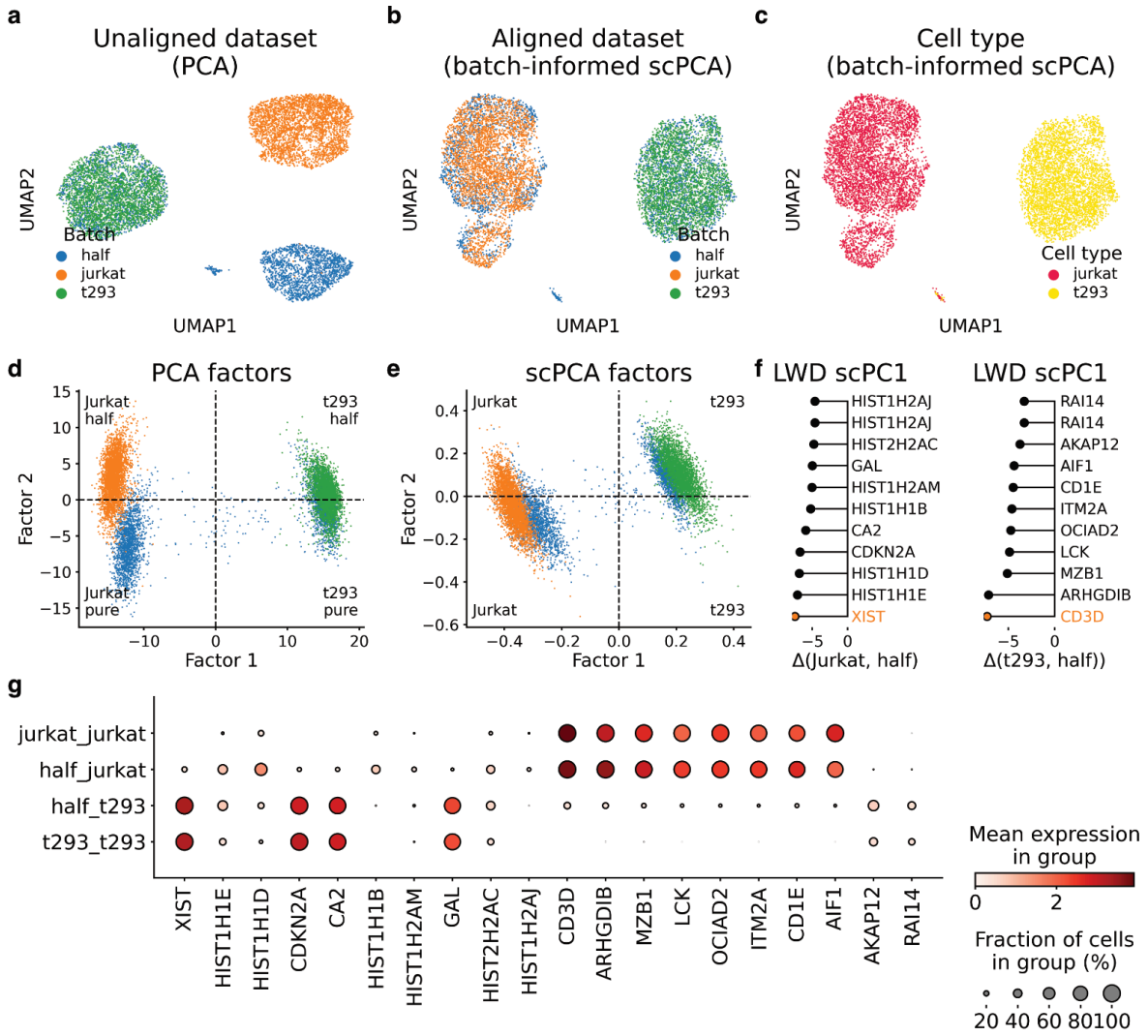


Figure 3: Modeling batch effects in cell-line data with scPCA. a. UMAP plot showing pure Jurkat and t293 cells, and a mixture of Jurkat and t293 cells (half) on the basis of a rank 20 PCA decomposition. **b. and c.** UMAP plots based on a rank 20 batch-informed scPCA decomposition with annotations for the batch (**b**) and cell line (**c**). **d. and e.** Scatter plots of factor 1 and 2 from the PCA (**d**) and the batch-informed scPCA (**e**) factorisation. **f.** Negative scPC1 loadings indicating genes that tend to be expressed at lower levels in the pure cell-line batches compared to the half mixture batch. **h.** Dot plot displaying the genes with the largest negative scPC1 LWDs across cell lines and batches.

To elucidate the genes responsible for the poor integration of Jurkat cells, we focused on scPC1, which delineated Jurkat from t293 cells with negative and positive factor weights, respectively (**Fig. 3e**). Analyzing the scPC1 LWDs between the pure Jurkat and the half-half mixture batch exposed genes with notably negative magnitude, suggesting high expression in cells of the half batch but not in pure Jurkat batch cells (**Fig. 3f**). Among these genes, we identified the t293 hallmark gene *XIST*, as well as *CDKN2A* and *CA2* which are highly expressed in t293 cells but not in Jurkat cells (**Fig. 3g**). Interestingly, scPCA detected histone encoding genes (*HIST1H1D*, *HIST1H1E* etc.) which were upregulated in Jurkat cells of the half batch but not in the pure Jurkat batch (**Fig. 3f**). Conversely, inspecting the LWDs

between the pure t293 and 50/50 mixture batch of scPC1 (**Fig. 3f**) indicated the down-regulation of Jurkat-specific genes, including its hallmark gene *CD3D* (**Fig. 3f, g**).

To assess scPCA's ability to model batch effects in more complex scenarios, we applied scPCA to a variety of benchmarking datasets (**Extended Data Fig. 1-6**)[18]. We observed that scPCA integrated data across different platforms (**Extended Data Fig. 1b, d**). Applying scPCA on a single-cell datasets of immune cells collected from multiple donors yielded an integrated data representation as visually confirmed through UMAP inspection (**Extended Data Fig. 2a**), but interestingly resulted in lower batch integration scores in comparison to baseline PCA (**Extended Data Fig. 2d**). More complex datasets, such as Luecken et al.'s lung and immune mouse dataset, which contained samples from different spatial locations, laboratories, protocols, and species, presented integration challenges (**Extended Data Fig. 3 and 4**). In these tasks, scPCA's linear framework was not sufficient to account for technical variation, a limitation that was visually evident through UMAP inspection. Finally, we tested scPCA on the two simulated single cell datasets, which were characterized by variations in cellular composition and nested batch effects, respectively (**Extended Data Fig. 5 and 6**). Here, we found that scPCA performed well on the simulated data set containing variable cellular composition, but encountered difficulties in integrating the dataset with nested batch effects.

In conclusion, our evaluation of scPCA across various benchmarking datasets demonstrated its robustness in handling batch effects tasks such as integrating data across platforms and handling datasets with multiple donors. However, scPCA's linear model exhibited limitations when addressing more complex batch covariates, such as spatial locations, laboratories, protocols, and species. These findings highlight the constraints of linear scRNA-seq integration methods in more intricate scenarios.

scPCA extracts cell type specific factors in complex sc-datasets

Previous findings suggest that scPCA's factor representation often aligns with cell types, and their corresponding components elucidate how their expression changes across conditions. To assess scPCA's utility in more complex single-cell data, we applied scPCA to lung samples of aging mice[33], which included more than 20 different cell types. This dataset contained 7,639 and 6,425 cells (total n=14,064) from mice sacrificed at 3 (3m) and 24 months (24m), respectively.

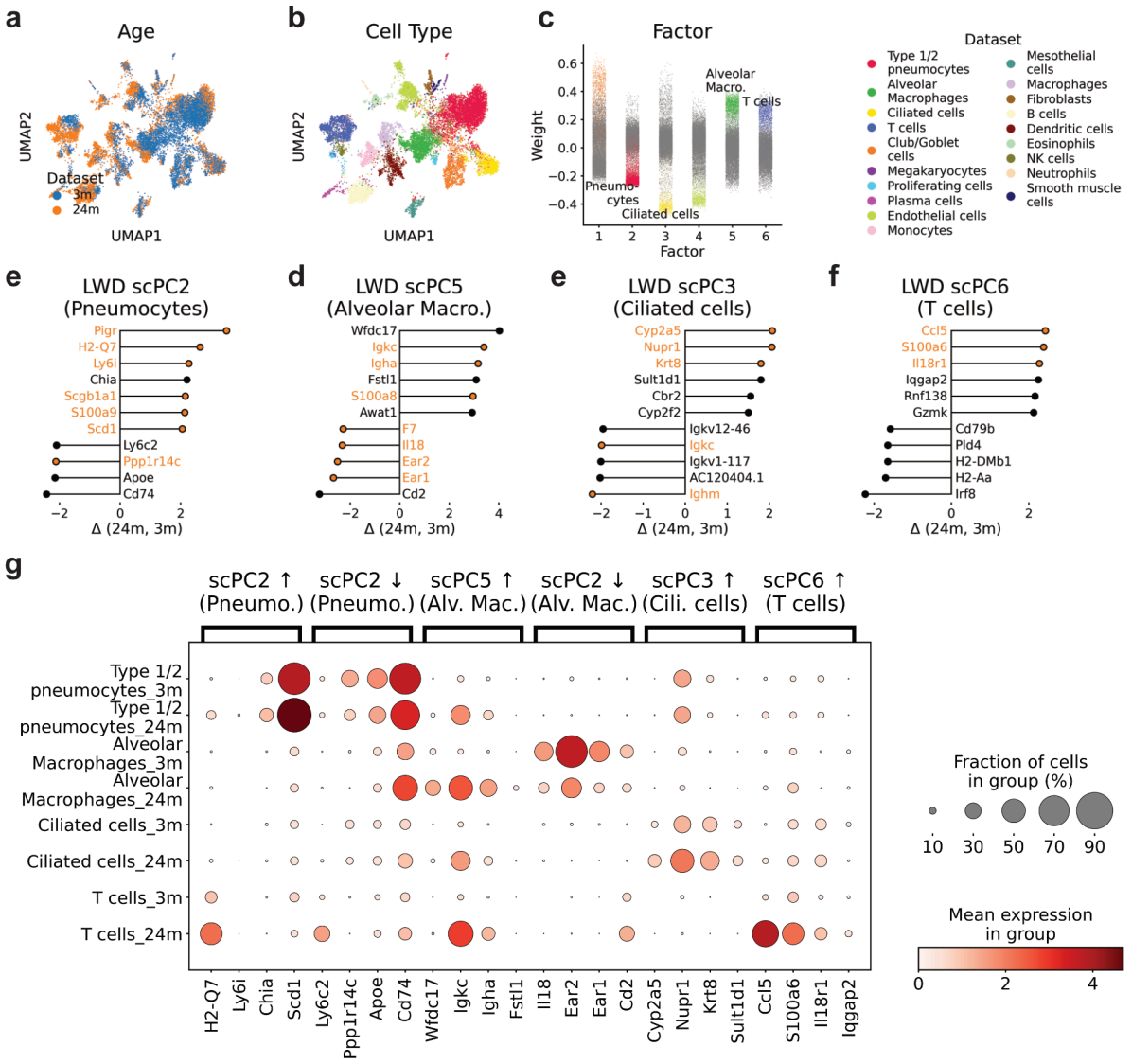


Figure 4: The scPCA decomposition sheds light on transcriptional changes in lung cells of aging mice. a and b. UMAP plots based on scPCA cell embeddings with annotations for the condition (**a**) and cell type (**b**). **c.** The scPCA factors show an enrichment for certain cell types. **e-f.** LWDs of scPC2, 5, 3 and 6 indicate genes that are either up or down regulated in aging mice. **g.** Dot plot displaying selected genes with the largest LWDs (**e-f**).

To account for the age of mice, we chose the 3m condition as the reference condition and modeled changes in cells of 24m old mice as deviations thereof (**Supplementary Fig. 4a**). Visual inspection of the scPCA-based UMAP indicated the integration of cells across conditions (**Fig. 4a, b**). We proceeded to analyze the factors, which revealed that the first factors showed an enrichment in various cell types (**Fig. 4c**). Particularly, the factors 2, 3, 5 and 6 displayed activity in pneumocytes, ciliated cells, alveolar macrophages and T cells, respectively (**Fig. 4c**).

Next, we subjected the top loading weights of scPC2, 3, 5 and 5 to GSEA, which revealed an enrichment of cell type-specific gene sets for each factor. For example, scPC2 was linked

to two processes related with surfactant metabolism, aligning with the enrichment of factor 2 in type II pneumocytes (**Supplementary Fig. 4b**). scPC3, with its factor predominantly enriched in ciliated cells, featured processes such as intraflagellar transport and cilium assembly (**Supplementary Fig. 4c**). As anticipated, factor 5 and 6 revealed gene sets associated with the immune system (**Supplementary Fig. 4d**), with the latter including pathways explicitly linked to T cells (**Supplementary Fig. 4e**).

To evaluate the impact of aging, we next computed LWDs across scPCs (**Fig. 4e-f**). Among the top upregulated genes of scPC2 were the MHC I associated gene *H2-Q7* and the Acyl-CoAl desaturase *Scd1* (**Fig. 4e, g**). Conversely, downregulated genes of scPC5 included *Ear1* and *Ear2* (**Fig. 4d, g**), corroborating the original authors' findings for type II pneumocytes and alveolar macrophages[33]. The examination of scPC3 LWDs, representative for ciliated cells, indicated an upregulation of genes associated with cellular stress response including *Nupr1* and *Cyp5a5* (**Fig. 4e, g**)[34,35]. Finally, changes along scPC5 indicated genes integral to immune response and regulation (**Fig. 4f, g**). These included *S100a6*, known for its role in cell stress[36,37]; *Ccl5*, a chemokine gene involved in inflammatory responses[38]; and *Il8r1*, which encodes a receptor important for immune cell migration[39].

We next applied GSEA to genes associated with the largest LWDs. This analysis revealed processes predominantly associated with oxidative stress and immune response (**Supplementary Fig. 4e-h**), aligning with the concept of ongoing 'inflammaging' in aging organisms[40]. For example, factor 3, enriched in ciliated cells, exhibited an enrichment of processes linked to oxidative stress (**Supplementary Fig. 4f**). One plausible explanation for this observation is that ciliated cells, positioned to face the lung lumen, are directly exposed to potentially harmful exogenous compounds.

In summary, we find scPCA's factor representation and its capability to model experimental states valuable for exploratory data analysis. In the context of single-cell data analysis, scPCA factors frequently align with cell-types, encapsulating their gene-expression programs within the associated components. Moreover, the linear parameterization of the loading tensor facilitates the interpretation of how individual scPCs change across conditions, thereby offering insights into either up or down-regulated gene programs, even in scenarios where the perturbation results in nuanced effect sizes.

Modeling multiple time points in scPCA

We then explored how scPCA disentangles single-cell datasets with conditioning variables including more than two levels. Particularly, we examined single-cell data from the visual cortex of mice that had been kept in darkness for seven days before being subjected to light exposure for durations of 0 (control), 1, and 4 hours (**Fig. 5a**)[41]. The dataset comprised a total of 48,266 single cells, distributed across the conditions with 16,188; 13,571; and 18,507 cells, respectively.

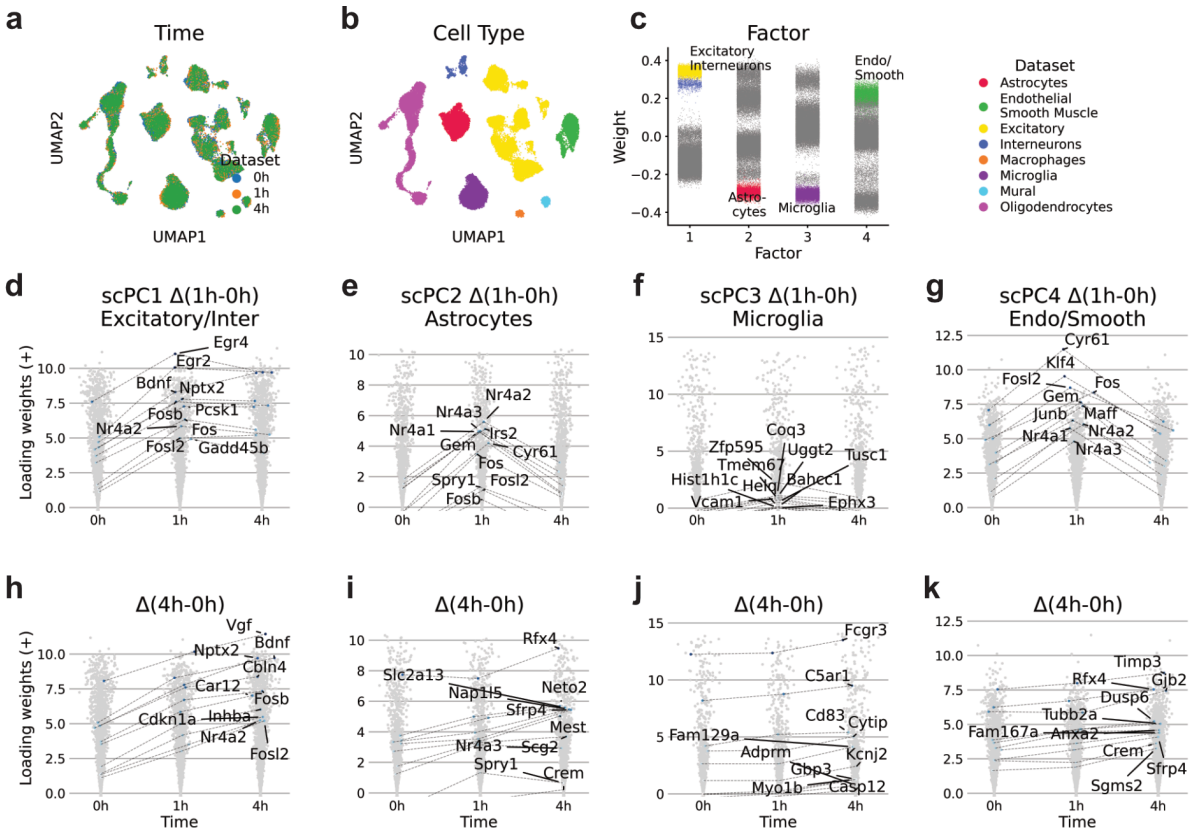


Figure 5: scPCA captures differentially expressed genes in light stimulated brain cells across multiple time points. a. and b. UMAP plots based on scPCA cell embeddings with annotations for the condition (**a**) and cell type (**b**). **c.** Factor weights of the first four scPCA components show an association with certain cell types. **d.-k.** Positive loading weights of scPC1-4 for the 0h, 1h and 4h time point. In **d.-g.** and **h.-k.**, the genes of the ten highest ranking LWDs for the 1h and 0h, and the 4h and 0h conditions are highlighted, respectively.

Exposure to depolarizing stimuli is thought to initiate two distinct phases of transcriptional changes. The first phase is controlled by early-response transcription factors (ERTFs), which are responsible for initiating the expression of late-response genes (LRGs)[42,43]. Interestingly, various types of neuronal cells follow a similar ERTF program, whereas the expression of LRGs, encompassing neuronal modulators and secreted factors, varies depending on the cell type.

We applied scPCA to the dataset by modeling the time points 1h and 4h with respect to the 0h reference condition (**Supplementary Figure 5a**). Subjecting the scPCA factors to UMAP dimensionality reduction indicated the integration across time points (**Fig. 5a, b**). A subsequent analysis of factors revealed that the factors 1, 2, 3, and 4 showed an enrichment in excitatory and interneurons, astrocytes, microglia, and endothelial and smooth muscle cells, respectively (**Fig. 5c**).

We then applied GSEA to the top loading weights of scPC1-4. This analysis revealed terms such as neuronal system or transmission across chemical synapses for scPC1, indicative for neuronal cells (**Supplementary Fig. 5b**). Likewise, the loading weights of scPC3 were indicative for gene sets such as “immune system”, consistent with the enrichment of factor 3 in microglia (**Supplementary Fig. 5d**). The GSEAs for scPC2 and 4 revealed gene

programs related to metabolism and compound transport, consistent with roles of astrocytes in metabolic support to neurons and endothelial cells in barrier and transport functions within the brain (**Supplementary Fig. 5c, e**).

Next, we computed LWDs for each scPC across the time points 0h and 1h. As anticipated, the examination of the top 10 LWDs of scPC1, enriched in excitatory and interneurons, revealed an upregulation of neuronal-specific ERTFs such as *Egr2*, *Egr4*, and *Fosb*, as well as common ERTFs like *Nr4a2* and *Fosl2* (**Fig. 5d**)[41]. Highest ranking LWDs of scPC2, associated with astrocytes, included non-neuronal ERTFs such as *Klf4*, alongside common ERTFs like *Nr4a1*, *Nr4a2*, *Nr4a3*, *Fos*, and *Fosl2* (**Fig. 5e**)[41]. Conversely, we detected only moderate loading weight changes and no ERTFs among scPC3 LWDs, consistent with the observations made by the authors for microglia (**Fig. 5f**). Finally, scPC4 LWDs displayed non-neural ERTFs like *Klf4* and *Maff*, and shared ERTFs including *Nr4a1*, *Nr4a2*, *Nr4a3*, *Fos*, and *Fosl2* (**Fig. 5g**).

We then subjected the top LWDs of scPC1-4 across the 1h and 0h condition to GSEA. Consistent with a shared ERTF response across cell types, this analysis revealed similar gene programs for scPC1, 2 and 4 LWDs, with the process "Signaling by NTRK1" being common to all of them (**Supplementary Fig. f, g, i**). NTRK1 is a receptor tyrosine kinase associated with the Ras/MAPK pathway, which is activated upon a depolarising stimulus. Beyond the set of shared terms, we observed that each component's gene sets also encompassed distinctive processes, which may mirror processes triggered by cell-type specific ERTFs. For example, scPC1 LDWs were indicative of signal transduction, whereas scPC4 gene sets were associated with pathways promoting cell survival and angiogenesis (**Supplementary Fig. f, i**).

We proceeded to investigate the longer ranging response to light stimulation by calculating LWDs between the 0h and 4h condition. Expectedly, the top 10 genes with largest positive LWDs of scPC1 contained the LRGs *Nptx2*, *Bdnf*, *Inhba*, *Cdkn1a*, *Cbln4*, *Car12* and *Vgf*[41], indicating that scPCA successfully captured stimulus-induced late transcriptional changes in excitatory and inhibitory neurons (**Fig. 5h**). Similarly, the top LDWs of scPC2, 3 and 4 included the LRG's *Rfx4*, *C5ar*, and *Anxa2*, respectively[41] (**Fig. 5i-k**).

Next, we applied GSEA to the top LWDs of scPC1-4 across the 4h and 0h condition. For scPC1, LWDs indicated ongoing neuronal activity (**Supplementary Fig. 5j**), which were characterized by gene sets that were also present in the GSEA for the 1h and 0h condition (**Supplementary Fig. 5f**). The LWDs for scPC2 and scPC4 were indicative of diverse intracellular processes (**Supplementary Fig. 5k, m**). In contrast to the shared gene sets seen in response to ERTFs (**Supplementary Fig. 5f-i**), these processes did not exhibit commonalities across cell types, suggesting that the response induced by LRGs is specific to each cell type. Interestingly, scPC3 LWDs displayed the upregulation of various gene sets connected to the innate immune system (**Supplementary Fig. 5l**). However, as the stimulus-dependent response in microglias was only modest, these results should be verified independently. In summary, we find that the scPCA's factorisation framework is capable of taking into account conditioning variables with multiple levels.

Modeling single-cell perturbation data with scPCA

In our final exploration, we aimed to assess whether scPCA effectively handles datasets with uneven cell type distributions across conditions (**Supplementary Fig. 6b**). To address this, we explored zebrafish data wherein the chordin locus was knocked out using CRISPR-Cas9 technology, resulting in zebrafish embryos with expanded ventral tissues[44]. This dataset contained 10,782 control (tyr) and 16,046 chordin knockout (chd) cells (total n=26,828).

We applied our algorithm by modeling chd cells with respect to the reference tyr condition (**Supplementary Fig. 6a**). Visual inspection of the scPCA-based UMAP indicated the integration of cells across conditions (**Fig. 6a**, **b**). We then proceeded to investigate the first three factors, which were enriched for epidermal, neural and tailbud cells, respectively (**Fig. 6c**).

We continued by conducting a GSEA on the top loading weights. For scPC1, this analysis included terms such as “embryonic cranial skeleton morphogenesis” and various other processes linked to the early formation of anatomical structures originating from epidermal cells (**Supplementary Fig. 6b**). The loading weights for scPC2 and scPC3 included gene sets associated with neuron differentiation and mesoderm development, respectively. This aligns with the enrichment of neural and tailbud cells in these factors (**Supplementary Fig. 6c, d**).

Next, we assessed the scPC1-3 LWDs to understand the transcriptional changes occurring along each scPC. Interestingly, this analysis indicated an upregulation of *szl* in the top LWDs of scPC1 and 3, but not in scPC2 (**Fig. 6d, f**). Chordin serves as an inhibitor of bone morphogenic protein (BMP), and *szl* operates as a ventral center molecule that is expressed under conditions of elevated BMP signaling[45]. Consequently, the absence of chordin in knockout cells could lead to an increase in *szl* transcript expression, a change that, to our knowledge, was not noticed in the author’s analysis of the data (**Supplementary Fig. 6i, j**). Additionally, we observed the upregulation of diverse differentiation factors, including *mdka*, *tbx2b*, *fsta*, *fgf8a*, *lbx2*, *fzd10*, and *hox10b* in chd cells, suggesting that the knockout induced widespread changes in the expressed differentiation programme of the developing zebrafish (**Fig. 6e, f, and g, Supplementary Fig. 6j**). Among the genes exhibiting negative LWDs, we found notch-dependent genes like *her1* and *her12* in scPC1 (**Fig. 6d, Supplementary Fig. 6j**), indicating that chordin loss affected broader regulatory networks that coordinate embryonic development.

Finally, we subjected scPC1-3 LWDs to GSEA. This analysis revealed gene ontology terms indicative for the development of organs originating from epidermal and neural cells for scPC1 and 2, respectively (**Supplementary Fig. 6f, g**). In line with the predominant activity of *chordin* in tailbud cells[44], the GSEA for scPC3 LWDs indicated the GO term “anterior/posterior specification”. Taken together, we find that scPCA effectively handles scRNA-seq datasets with differences in cell type distribution across conditions.

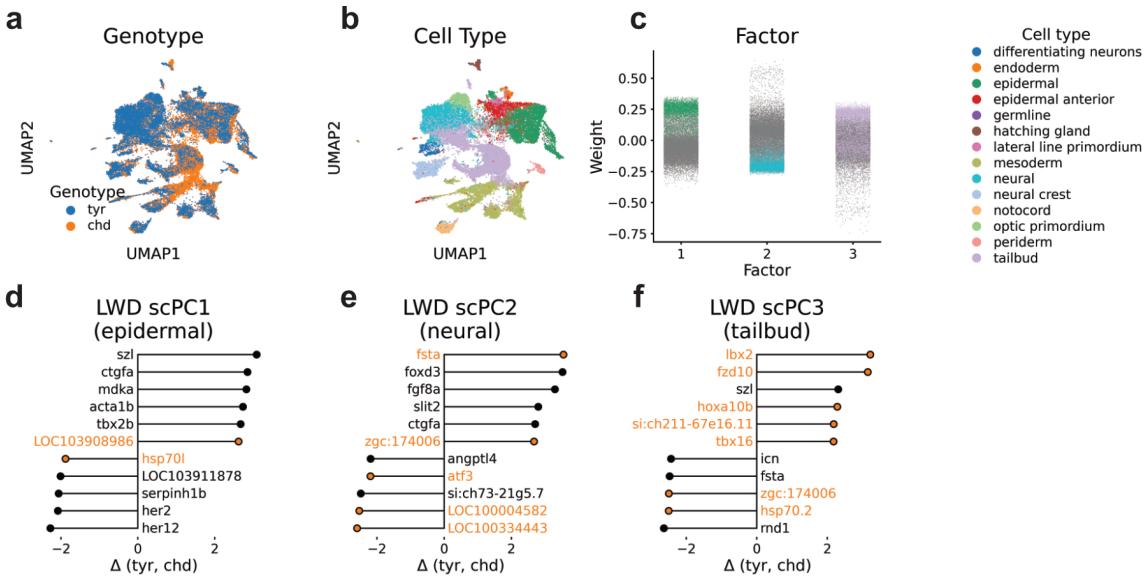


Figure 6: Modeling single-cell perturbation data with scPCA. a. and b. UMAP plots based on scPCA cell embeddings with annotations for the perturbation (**a**) and cell type (**b**). **c.** Factor weights of scPCA components 1-3. Each dot represents a cell, colors indicate the cell-type. **d.-f.** Highest ranking LWDs of scPC1-3 indicate genes that are either up or down regulated across conditions.

Discussion

We introduced scPCA, a factorization model designed to analyze multi-condition single-cell data. By informing our factor model about the conditioning variable, we define components for each condition relative to a chosen reference. This approach enables the model to establish distinct but related coordinate systems, effectively addressing shifts in gene expression patterns caused by experimental interventions. As a result, scPCA generates integrated cell embeddings and provides interpretable gene loading weights. Moreover, by calculating the difference in gene loading weights across conditions, the model highlights factor-specific patterns of upregulated or downregulated gene expression.

We applied our method across a range of single-cell datasets, showcasing the effectiveness of scPCA in handling a variety of scenarios. This included modeling both stark and modest perturbation effects, multiple time points, datasets characterized by differences in cell type distribution across conditions, and integrating batch effects. We anticipate that the linear model of scPCA will be instrumental in integrating and interpreting multi-condition single-cell datasets. This approach holds promise for enhancing our understanding of intricate biological processes captured in such data.

Despite its strengths, scPCA has some limitations. A key constraint is its assumption that the principal axes of variation are largely conserved across conditions. Because scPCA expresses the bases of each condition as a linear function of the reference basis/condition, the model primarily captures linear expression changes. This means it may struggle in scenarios where gene expression changes are non-linear. Additionally, scPCA is best positioned as a tool for exploratory analysis. While loading weight differences often highlight up- or down-regulated genes, these findings should be further validated through

complementary methods. Lastly, our approach uses approximate Bayesian inference to estimate the posterior distribution of parameter weights, which can lead to an underestimation of their variance, potentially affecting the precision of these estimates.

Methods

The scPCA model

For a complete derivation of the model see the Supplementary Methods. The scPCA factorisation requires a multi-condition single-cell expression count matrix $\mathbf{X} \in \mathbb{N}_0^{C \times G}$ for C cells and G genes, a design matrix $\mathbf{D} \in \{0,1\}^{C \times Q}$ that encodes a conditioning variable with Q levels, and an optional indicator matrix $\mathbf{B} \in \{0,1\}^{C \times P}$ that may adjust for condition-specific mean gene expression for P groups of cells ($P = 1$ or $P = Q$, i.e. scPCA fits a global or condition-specific mean offset(s) respectively) as inputs. We model the elements of the expression count matrix $\mathbf{X} = \{x_{cg}\}$ for cell $c = 1, \ldots, C$ and gene $g = 1, \ldots, G$ as negative binomial distributed, given an unobserved gene expression level μ_{cg} and a gene-specific over-dispersion α_g

$$x_{cg} \sim \mathrm{NB}(\mu_{cg}, \alpha_g). \tag{1}$$

In scPCA, we decompose the unobserved gene expression as a linear function of the conditioning variable encoded in $\mathbf{D}$. Denote the entries of the design and indicator matrix $\mathbf{D} = \{d_{cq}\}$ and $\mathbf{B} = \{b_{cp}\}$ for level $q = 1, \ldots, Q$ and group $p = 1, \ldots, P$ respectively, and let $l_c = \log \sum_{g=1}^{G} x_{cg}$ denote a cell-specific fixed size factor (total number of counts). Then,

$$\log \mu_{cg} = \underbrace{l_c}_{\text{Size factor}} + \underbrace{\left(\sum_{p=1}^{P} b_{cp} v_{pg} \right)_{cg}}_{\text{Intercept term}} + \underbrace{\left(\sum_{q=1}^{Q} d_{cq} \cdot \sum_{f=1}^{F} z_{cf} w_{fgq} \right)_{cg}}_{\text{Factor model}} \tag{2}$$

where w_{fgq} represents the entries of a loading weight tensor $\mathbf{W} \in \mathbb{R}^{F \times G \times Q}$ consisting of $f = 1, \ldots, F$ components, z_{cf} the elements a factor weight matrix $\mathbf{Z} \in \mathbb{R}^{C \times F}$ containing the coordinates of each cell with respect to $\mathbf{W}$, and v_{pg} the entries of a mean offset $\mathbf{V} \in \mathbb{R}^{P \times G}$ matrix.

Interpretation of the factors **Z**

Similar to interpreting factors in PCA, scPCA factors denote lower-dimensional representations of global sources of variability in the data. Each factor arranges cells along a principal axis of variation centered at zero. Cells with positive or negative factor weights manifest opposite phenotypes, with the magnitude of the factor weight indicating the strength of the effect.

Interpretation of the loading weights **W**

The loading weights provide a score for the importance of each gene to its corresponding factor. Loading weights close to zero indicate genes with little or no importance to the factor, whereas genes with large absolute values exhibit a strong factor association. The direction of the effect is indicated by the sign of the weight: a positive weight suggests higher gene expression in cells with positive factor values, and conversely for negative weights.

Central to scPCA's modeling approach is its ability to infer loading weight differences (LWDs) across the levels of the conditioning variables. This enables an assessment of gene expression shifts along each component of the decomposition. To accomplish this, scPCA relies on the user to establish a reference level for comparison, and encodes the categorical conditioning variables in terms of a design matrix. For example, in the common two-class comparison setting (e.g., control versus treated), $\mathbf{D}$ is a $C \times 2$ design matrix with a column of ones corresponding to an intercept and a column of indicator variables for the class of each cell (e.g., 0 for control and 1 for treated). Here, the intercept column defines the reference level, and instructs the model to fit the basis of the treatment condition as an offset of the basis of the control condition (for detailed derivation please refer to the Supplementary Methods).

Note on the mean offset parameter **V**

In scPCA, we directly model single-cell expression data in count space, thereby evading the need to transform the data prior to the analysis. However, this prohibits to mean-center the data which is an obligatory preprocessing step for PCA. For this reason, we introduce an explicit parameter to capture either the global or condition-specific mean expression of each gene. Empirically, we observed that for most single-cell datasets a global mean offset is sufficient (in this case B reduces to a vector of ones). However, additional adjustments may be necessary in cases when the conditioning variable affects expression globally.

Model inference

We employ stochastic variational inference to estimate the model parameters and implemented scPCA in the probabilistic programming language Pyro[23], which supports GPU acceleration. Note that scPCA produces probabilistic results, meaning parameter estimates may vary slightly with different algorithm initializations. To ensure optimal results, we recommend users run multiple initializations of the model and select the one with the highest ELBO score. For full details, see the Supplementary Methods.

Model selection

Determining the appropriate number of factors is challenging: too few factors might miss important biological insights, while too many factors can cause an "oversegmentation" of the data, thus complicating interpretation. In our experiments, we found that for most mid-sized single-cell datasets, 20 factors provided a good balance between complexity and interpretability. We note that scPCA returns its factors by importance, i.e. ranked by the variance explained of each individual factor.

Simulation Experiments

For the simulation experiments, we simulated data according to the generative model (Supplementary Methods) using $F = 10$ factors, $G = 2000$ genes and $Q = 2$ conditions. To assess the scPCA inference under various settings, we varied the number of cells per condition $C_q \in \{10, 100, 1000, 10000\}$ and used different levels of negative binomial noise to simulate the data.

Subspace reconstruction

Recall that different sets of basis vectors $\{\mathbf{w}_1, \ldots, \mathbf{w}_F\}$ and $\{\mathbf{v}_1, \ldots, \mathbf{v}_F\}$ may span the same vector space. More formally, if $\mathbf{W}$ and $\mathbf{V}$ consist of the column vectors $\{\mathbf{w}_1, \ldots, \mathbf{w}_F\}$ and $\{\mathbf{v}_1, \ldots, \mathbf{v}_F\}$ respectively and $\mathbf{x} = \mathbf{W}\mathbf{y}$, then there exists a vector $\mathbf{z}$ such that $\mathbf{x} = \mathbf{V}\mathbf{z}$. To see this, consider first the case when $\mathbf{W}$ and $\mathbf{V}$ are full rank, then $\mathbf{W}$ and $\mathbf{V}$ are both invertible. Then

$$\mathbf{x} = \mathbf{W}\mathbf{y} = \mathbf{V}\mathbf{V}^{-1}\mathbf{W}\mathbf{y} = \mathbf{V}\mathbf{z}$$

where $\mathbf{z} = \mathbf{V}^{-1}\mathbf{W}\mathbf{y}$.

However, since scPCA decompositions are generally not full rank (i.e. $F << G$) we instead consider projection matrix of the estimated loading tensor $\widehat{\mathbf{W}}$,

$$\mathbf{P} = \widehat{\mathbf{W}}(\widehat{\mathbf{W}}^T\widehat{\mathbf{W}})^{-1}\widehat{\mathbf{W}}^T.$$

If $\widehat{\mathbf{W}}$ and the simulated ground truth $\mathbf{W}$ span the same subspace, then $\mathbf{P}\mathbf{W} = \mathbf{W}$. Thus, in order to evaluate how accurate scPCA reconstructs the data-generating subspaces, we compute the root mean squared error (RMSE) between $\mathbf{P}\mathbf{W}$ and $\mathbf{W}$.

Model selection

To find the appropriate decomposition rank for a given dataset, we propose to employ the root means square error (RMSE). In our simulation experiments, we found a sharp flattening of the RMSE at the true rank of the data (Supplementary Fig. 1a).

Single-cell data preprocessing and analysis

We downloaded the filtered raw-count single-cell data from Zenodo[17], GSE124872[33], GSE102827[41], 10xGenomics[32], Figshare[18] and GSE112294 [44]. We filtered genes using highly variable gene selection (HVG) using scanpy's "seurat_v3" setting to reduce the number of total genes to 4000. We then applied the logCPM10k transformation to the count data before subjecting them to a rank 20 PCA decomposition. For the scPCA analysis, we performed an equivalent rank 20 decomposition on the raw count data.

Gene Set Enrichment Analysis

To associate scPCA factors with biological processes, we used the gseapy Python package (version 1.0.4) using either Reactome or Gene Ontology genests. We used the top 200-300

genes with highest or lowest loadings and performed gene set overrepresentation analysis using all genes in the data set as background.

Code and Data availability

The scPCA software package is available on pypi. The source code and documentation are available on Github. The datasets used in this work may be downloaded from Zenodo[17], GSE124872[33], GSE102827[41], 10xGenomics[32], Figshare[18] and GSE112294 [44].

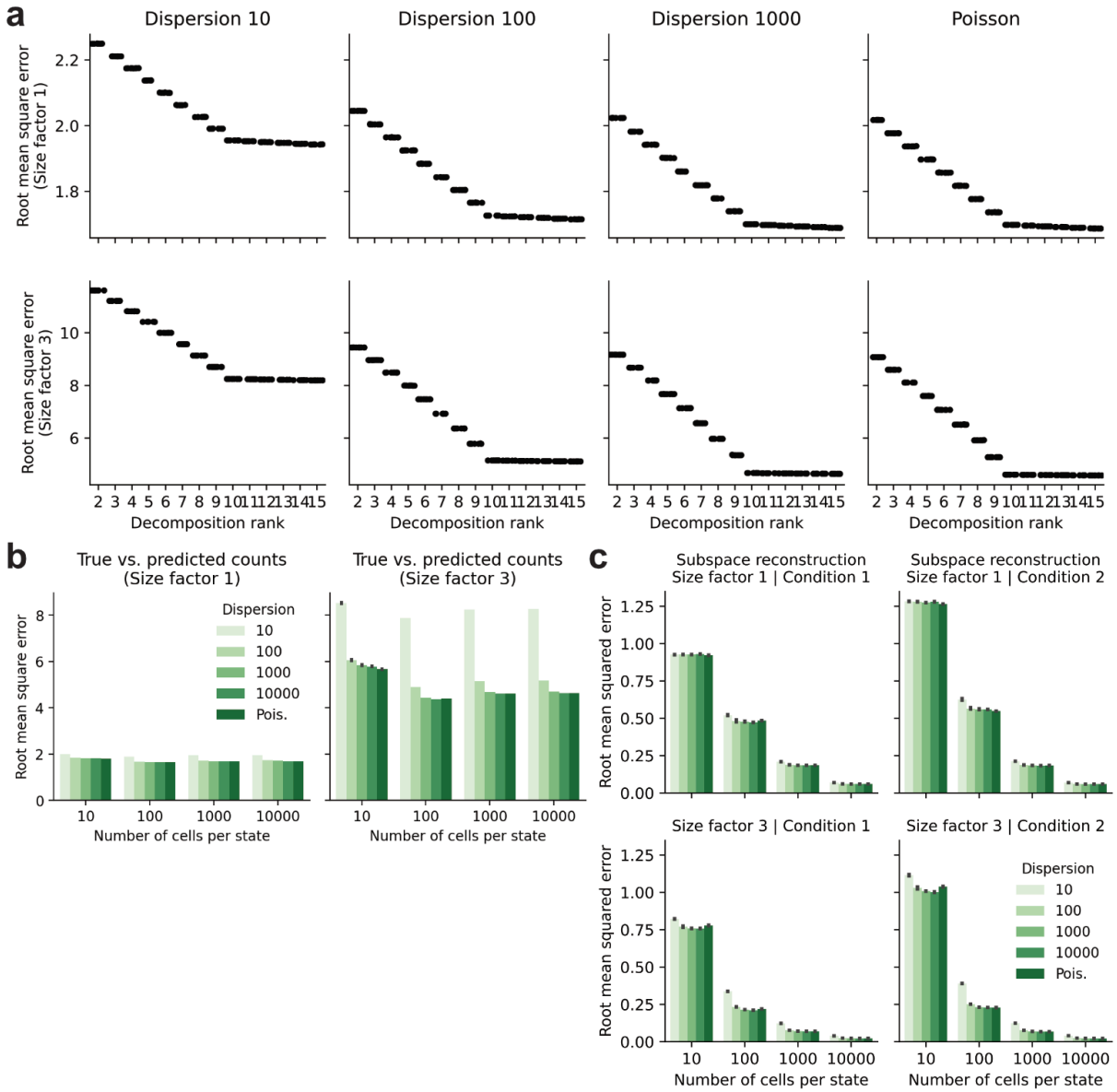
Acknowledgements

We would like to thank Wolfgang Huber, Tumay Capraz, Stefan Pleidli, Constantin Ahlman-Eltze and Petr Smirnov for useful discussions and comments on the paper. We also thank Britta Velten, Martin Rohbeck and Kai Ueltzhöffer for insightful comments on the manuscript.

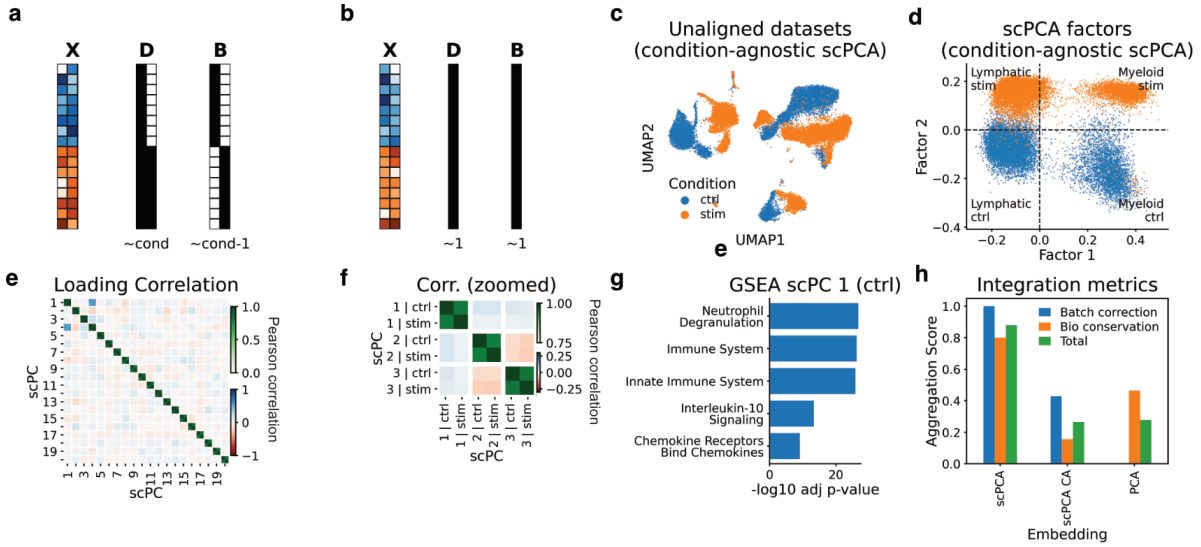
Author contributions

H.V. conceived and implemented the scPCA software. H.V wrote the paper and designed the figures with input from S.D. S.D. supervised the work.

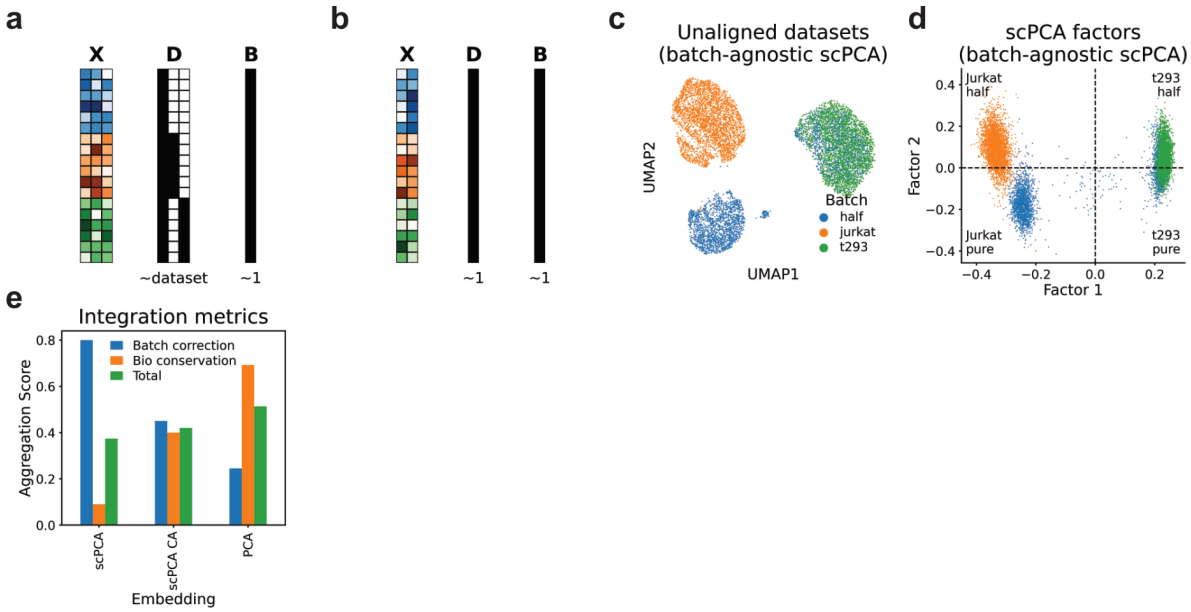
Supplementary Figures



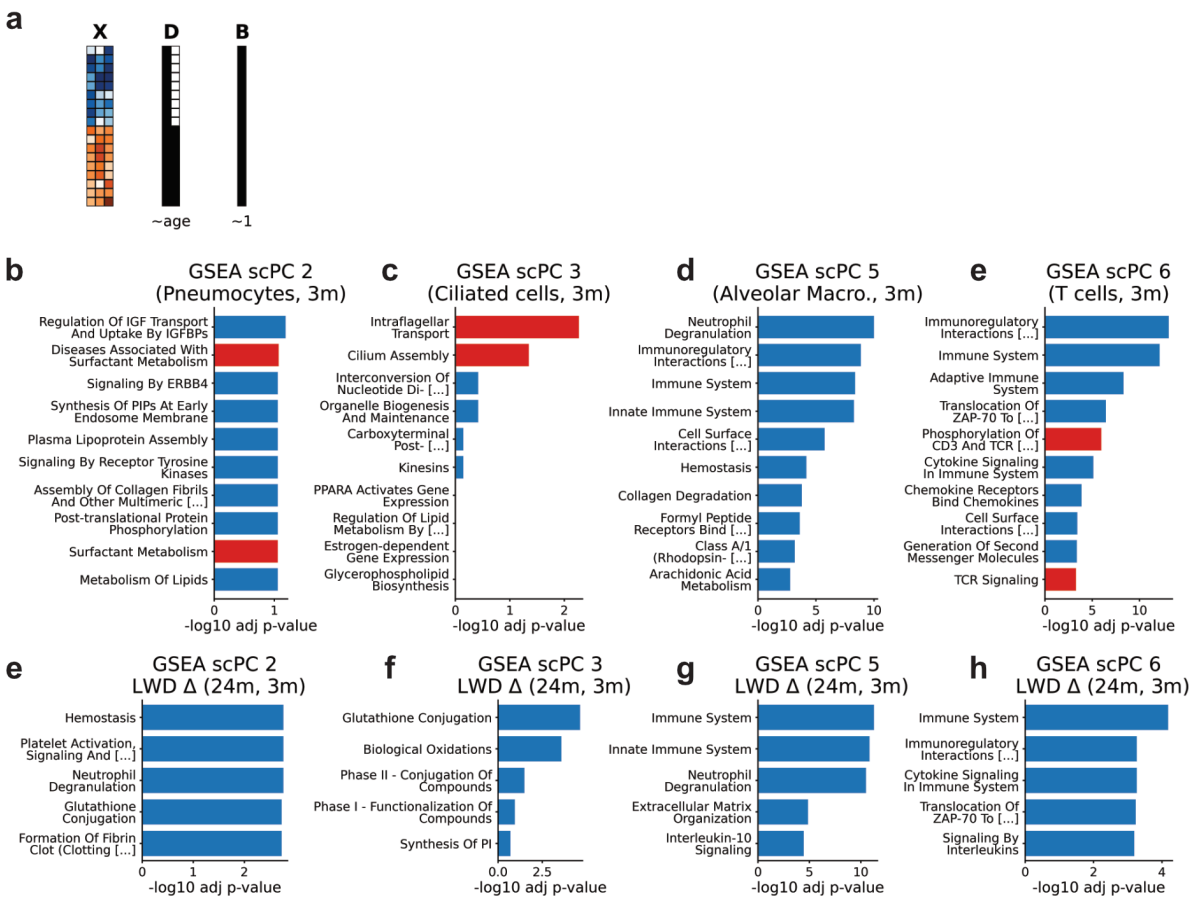
Supplementary Figure 1: Simulation results For **a.**, **b.** and **c.** In-silico data comprising two conditions were simulated according to the generative process of the model (**Methods**) to assess how well the scPCA inference reconstructs simulated ground truth. The parameters of these simulations were chosen to cover a range of noise levels (overdispersion), average number of counts per cell (size factor) and number of cells per condition. **a.** Model selection via the root mean squared error (RMSE) of true and predicted counts. Data was simulated using ten basis vectors ($F = 10$), two conditions $Q = 2$ and $C_q = 1000$ cells per condition. For each parameter combination, we performed 10 scPCA decompositions for the ranks 2 to 15. **b.** RMSE of true and predicted counts for dataset of different sizes (total number of cells) and number of counts per cell (size factor). **c.** RMSE of the true and predicted subspace of the first and second simulated condition of the data.



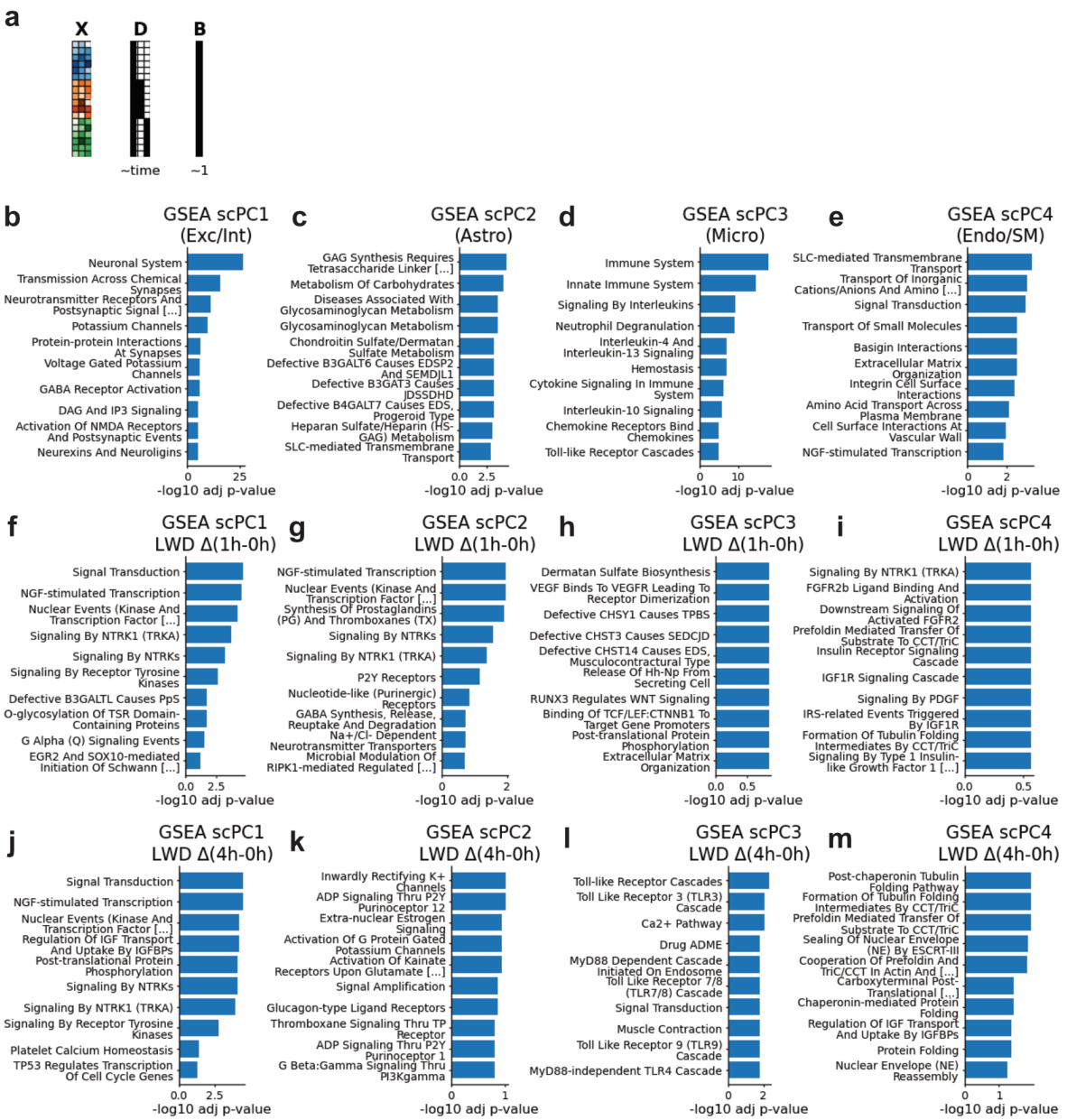
Supplementary Figure 2: Modeling untreated and IFN beta stimulated PBMCs. a. and b. Schematic depiction of the input data for the condition-informed (**a**) and condition-agnostic (**b**) scPCA factorizations. These include the expression count matrix **X** with cells from the control (blue) and stimulated (orange) condition, alongside the design and indicator matrices **D** and **B**. The design matrix **D** uses the control condition as a reference (column of ones) and models the IFN-beta condition as a deviation from this baseline (**a**). The indicator matrix **B** is used to fit independent mean offsets for each condition (**a**). Conversely, using columns of ones as design matrices in scPCA results in dimensionality reduction that is agnostic of condition differences (**b**). **c.** UMAP plot based on the condition-agnostic scPCA factorization. **d.** Scatter plots of the first two scPCA factors (condition-agnostic). **e. and f.** The heatmap of loading weight cross correlations indicate low correlations between different components (red/blue colorbar), yet high correlations between the same component across conditions (green colorbar). **g.** Gene set enrichment analysis based on the top 200 loading weights of scPC1 (condition-informed). **h.** Comparison of batch integration and bio-conservations scores for cell embeddings generated by condition-aware and agnostic (CA) scPCA models as well as conventional PCA.



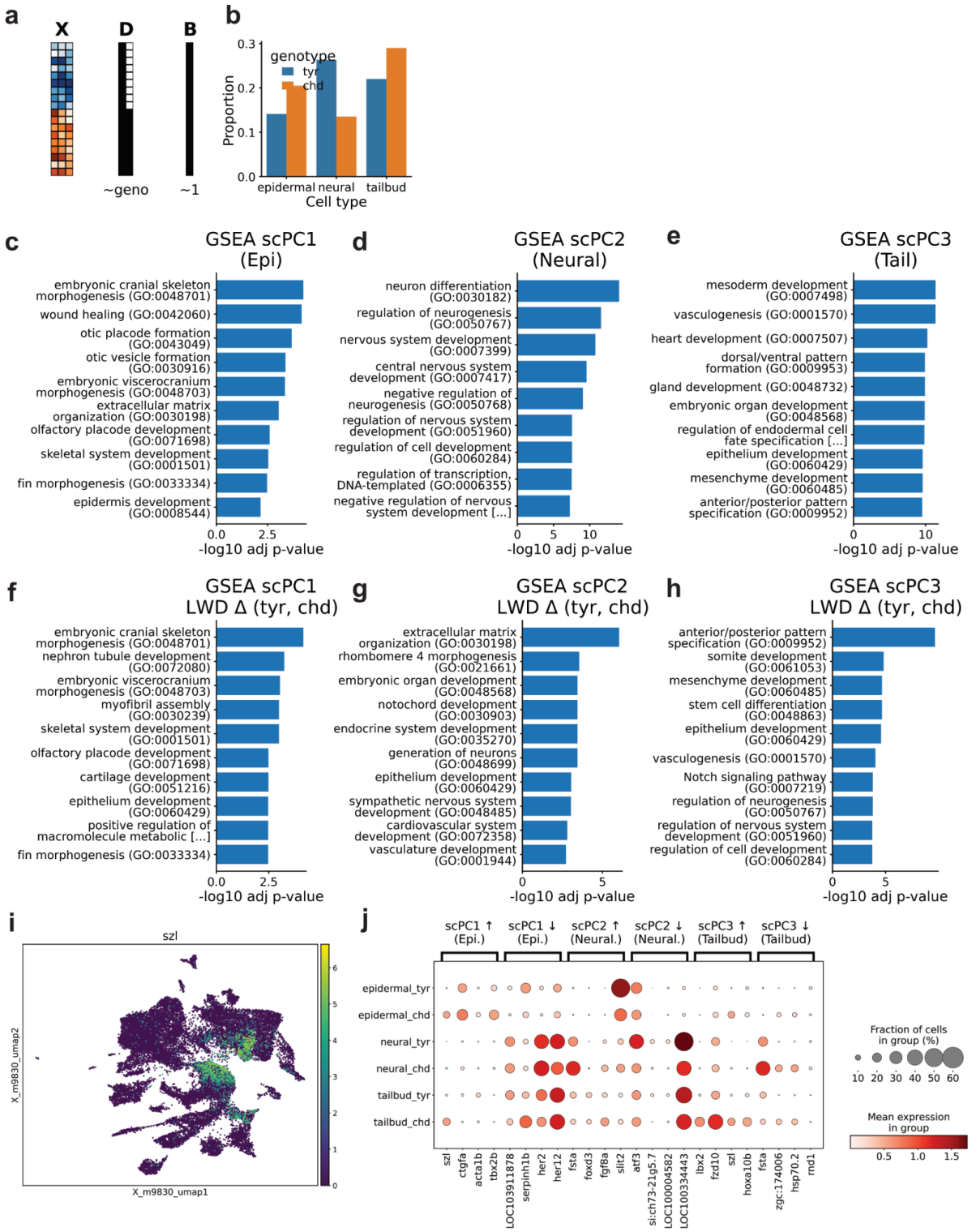
Supplementary Figure 3: a. and b. Schematic depiction of the input data for the batch-informed (**a**) and batch-agnostic (**b**) scPCA factorizations, including the expression count matrix **X** with cells from the half-half batch (blue), pure Jurkat batch (orange), and pure t293 batch (green), alongside the design and intercept matrices **D** and **B**. In (**a**), the loading design matrix **D** sets the half-half batch as a reference (column of ones) and aligns the components for the Jurkat and t293 batches, respectively. An indicator matrix of ones allows the model to fit a single mean offset across all batches. Conversely, using columns of ones for both design and indicator matrices leads to dimensionality reduction that ignores the batch structure in the data (**b**). **c.** The UMAP plot of Jurkat and t293 cells on the basis of the batch-agnostic scPCA factorisation results in a representation in which Jurkat cells separate by batch. **d.** Scatter plot of the first two factors from the batch-agnostic scPCA factorization, highlighting how factor 2 distinctly separates Jurkat cells by batch. **e.** Comparison of batch integration and bio-conservations scores for cell embeddings generated by the batch-aware and agnostic scPCA models as well as conventional PCA.



Supplementary Fig. 4: a. Schematic depiction of the input data including the expression count matrix **X** with lung cells from the 3 (blue) and 24 (orange) month old mice, alongside the design and intercept matrices **D** and **B**. **b.-e.** GSEA based on the top loading weights of indicated scPCs. Highlighted gene sets indicate particularly relevant programs (see main text). **f.-i.** GSEA based on positive LWDs of scPC2, 3, 5 and 6 indicate upregulated gene programs in aging mice.



Supplementary Figure 5: a. Schematic depiction of the input data for scPCA including the expression count matrix **X** with mouse cortex cells from the timepoints 0 (blue), 1 (orange) and 3 hours, alongside the design and intercept matrices **D** and **B**. **b.-e.** GSEA based on the top loading weights of indicated scPCs. Highlighted gene sets indicate particularly relevant programs. **f.-i.** GSEA based on positive LWDs of scPC1-4 indicate upregulated gene programs across the 1h and 0h timepoints. **j.-m.** GSEA based on positive LWDs of scPC1-4 indicate upregulated gene programs across the 5h and 0h timepoints.



Supplementary Figure 6: a. Schematic depiction of the input data for scPCA including the expression count matrix **X** with zebrafish cells from the control (blue) and chordin knockout (orange), alongside the design and intercept matrices **D** and **B**. **c.-e.** GSEA based on the top loading weights of indicated scPCs. **f.-h.** GSEAs based on positive LWDs of scPC1-3. **i.** UMAP plot showing the expression of *szl* in distinct epidermal and tailbud cells. **j.** Dot plot representation of selected genes exhibiting the largest LWDs in indicated scPCs.

References

1. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
2. Kiselev, V. Y., Andrews, T. S. & Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.* **20**, 273–282 (2019).
3. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018).
4. Maaten, L. & Hinton, G. E. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
5. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
6. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
7. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).
8. Madrigal, A., Lu, T., Soto, L. M. & Najafabadi, H. S. A unified model for interpretable latent embedding of multi-sample, multi-condition single-cell data. *bioRxiv* 2023.08.15.553327 (2023) doi:10.1101/2023.08.15.553327.
9. Abid, A., Zhang, M. J., Bagaria, V. K. & Zou, J. Contrastive Principal Component Analysis. *arXiv [stat.ML]* (2017).
10. Li, D., Jones, A. & Engelhardt, B. Probabilistic Contrastive Principal Component Analysis. *arXiv [stat.ME]* (2020).
11. Argelaguet, R. *et al.* MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, 111 (2020).
12. Sohn, K., Lee, H. & Yan, X. Learning structured output representation using deep

conditional generative models. *Adv. Neural Inf. Process. Syst.* 3483–3491 (2015).

13. Lotfollahi, M., Naghipourfar, M., Theis, F. J. & Wolf, F. A. Conditional out-of-distribution generation for unpaired data using transfer VAE. *Bioinformatics* **36**, i610–i617 (2020).
14. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nat. Methods* **16**, 715–721 (2019).
15. 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).
16. Gayoso, A. *et al.* Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat. Methods* **18**, 272–282 (2021).
17. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2017).
18. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
19. Dixit, A. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853–1866.e17 (2016).
20. Datlinger, P. *et al.* Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297–301 (2017).
21. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).
22. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
23. Bingham, E. *et al.* Pyro: Deep Universal Probabilistic Programming. *arXiv [cs.LG]* (2018).
24. Hoffman, M. D., Blei, D. M., Wang, C. & Paisley, J. Stochastic Variational Inference. *J. Mach. Learn. Res.* **14**, 1303–1347 (2013).
25. Ranganath, R., Gerrish, S. & Blei, D. M. Black Box Variational Inference. *arXiv [stat.ML]* (2013).
26. 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).

27. Svensson, V. Droplet scRNA-seq is not zero-inflated. *Nat. Biotechnol.* **38**, 147–150 (2020).
28. Fernbach, S. *et al.* Restriction factor screening identifies RABGAP1L-mediated disruption of endocytosis as a host antiviral defense. *Cell Rep.* **38**, 110549 (2022).
29. Tran, H. T. N. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* **21**, 12 (2020).
30. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21 (2019).
31. Shaham, U. *et al.* Removal of batch effects using distribution-matching residual networks. *Bioinformatics* **33**, 2539–2546 (2017).
32. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
33. Angelidis, I. *et al.* An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nat. Commun.* **10**, 963 (2019).
34. Cano, C. E., Hamidi, T., Sandi, M. J. & Iovanna, J. L. Nupr1: the Swiss-knife of cancer. *J. Cell. Physiol.* **226**, 1439–1443 (2011).
35. Huang, C., Santofimia-Castaño, P. & Iovanna, J. NUPR1: A Critical Regulator of the Antioxidant System. *Cancers* **13**, (2021).
36. Foell, D., Wittkowski, H., Vogl, T. & Roth, J. S100 proteins expressed in phagocytes: a novel group of damage-associated molecular pattern molecules. *J. Leukoc. Biol.* **81**, 28–37 (2007).
37. Wang, Y., Kang, X., Kang, X. & Yang, F. S100A6: molecular function and biomarker role. *Biomark Res* **11**, 78 (2023).
38. Huffman, A. P., Lin, J. H., Kim, S. I., Byrne, K. T. & Vonderheide, R. H. CCL5 mediates CD40-driven CD4+ T cell tumor infiltration and immunity. *JCI Insight* **5**, (2020).
39. Jones, S. A., Wolf, M., Qin, S., Mackay, C. R. & Baggiolini, M. Different functions for the

interleukin 8 receptors (IL-8R) of human neutrophil leukocytes: NADPH oxidase and phospholipase D are activated through IL-8R1 but not IL-8R2. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 6682–6686 (1996).

40. Panda, A. *et al.* Human innate immunosenescence: causes and consequences for immunity in old age. *Trends Immunol.* **30**, 325–333 (2009).
41. Hrvatin, S. *et al.* Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. *Nat. Neurosci.* **21**, 120–129 (2018).
42. Harward, S. C. *et al.* Autocrine BDNF-TrkB signalling within a single dendritic spine. *Nature* **538**, 99–103 (2016).
43. Mardinly, A. R. *et al.* Sensory experience regulates cortical inhibition by inducing IGF1 in VIP neurons. *Nature* **531**, 371–375 (2016).
44. Wagner, D. E. *et al.* Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).
45. De Robertis, E. M. & Moriyama, Y. Chapter Thirteen - The Chordin Morphogenetic Pathway. in *Current Topics in Developmental Biology* (ed. Wassarman, P. M.) vol. 116 231–245 (Academic Press, 2016).