METHOD

eSVD-DE: Cohort-wide differential expression in single-cell RNA-seq data using exponential-family embeddings

Kevin Z. Lin[1*], Yixuan Qiu[2] and Kathryn Roeder[3]

[*]Correspondence: kzlin@uw.edu
[1]Department of Biostatistics, University of Washington, Seattle, USA
Full list of author information is available at the end of the article

Abstract

Single-cell RNA-sequencing (scRNA) datasets are becoming increasingly popular in clinical and cohort studies, but there is a lack of methods to investigate differentially expressed (DE) genes among such datasets with numerous individuals. While numerous methods exist to find DE genes for scRNA data from limited individuals, differential-expression testing for large cohorts of case and control individuals using scRNA data poses unique challenges due to substantial effects of human variation, i.e., individual-level confounding covariates that are difficult to account for in the presence of sparsely-observed genes. In this paper, we develop the eSVD-DE, a matrix factorization that pools information across genes and removes confounding covariate effects, followed by a novel two-sample test in mean expression between case and control individuals. In general, differential testing after dimension reduction yields an inflation of Type-1 errors. However, we overcome this by testing for differences between the case and control individuals' posterior mean distributions via a hierarchical model. We benchmark our method's accuracy and power increase compared to other DE methods typically repurposed for analyzing cohort-wide differential expression based on simulated data. We then use the eSVD-DE to study the cohort-wide differential expression in idiopathic pulmonary fibrosis, varying severity of ulcerative colitis, and autism spectrum disorder. Altogether, eSVD-DE proposes a novel and powerful way to test for DE genes among cohorts after performing a dimension reduction.

Keywords: case-control subjects; Gamma-Poisson distribution; matrix factorization; multi-individual data

Background

High-throughput single-cell RNA-seq (scRNA) technology has advanced tremendously over the last decade and helped biologists uncover differing cell-type proportions as well as differentially expressed (DE) genes within a particular cell-type when studying various diseases or disorders. These findings were previously inaccessible using bulk RNA-seq technology. As the technology has developed more in accuracy and cost-efficiency, many labs have begun sequencing entire cohorts of individuals to study up-regulated/down-regulated pathways specific to each cell type in diseases/disorders that generalize to the human population. For example, biologists have sequenced hundreds of thousands of cells for 44 individuals with lung adenocarcinoma [1], hundreds of thousands of neurons for 84 individuals with varying severity of Alzheimer's Disease [2], and millions of blood cells for 162 indi-

viduals with systemic lupus erythematosus and 99 control individuals [3] in order to discover DE genes for the disease/disorder consistent with the entire cohort. While numerous existing and benchmarked single-cell DE methods are designed to find differentially-expressed patterns among *cells* [4, 5], this task fundamentally differs from the cohort-wide studies' goal of finding differentially expressed patterns among *individuals*. This pressing methodological gap has been raised by many biologists who have collected cohort-wide scRNA-seq datasets [6, 7]. Hence, we reinvestigate the shortcomings of existing DE methods commonly used to analyze cohort-wide scRNA-seq data and design a new DE method specifically suited for such data. Focusing on scRNA-seq data enables us to study more principled ways to model cohort-wide scRNA-seq data, which we hope could be used to inspire cohort-wide DE methods for other single-cell technologies in the future, as well as pioneer applications in future eQTL, as well as upcoming studies that sequence spatial transcriptomics or paired multiomics on large cohorts.

One of the main distinctions between a DE analysis among cells compared to a DE analysis among individuals lies in how the case and control population distributions are quantified. DE analyses among individuals need to account for variability within and among individuals in order to properly model human variation. However, most existing DE methods lack one of these two aspects. On the one hand, variability within an individual hinders "pseudobulk" analyses where all the cells among each individual are summed to yield a pseudobulk sample. Then, methods originally designed for bulk RNA-seq data such as DESeq2 [8] and edgeR [9] are used, but these methods do not account for the variability within an individual. On the other hand, variability across individuals hinders most DE methods made for scRNA-seq data. Specifically, a gene could be differentially expressed among cells but not among individuals if all cells with a significantly higher gene expression come from a small subset of individuals.

The second main distinction is that in cohort-wide scRNA-seq data, there are potentially substantial effects of individual-level covariates such as age, sex, and smoking status that can induce differences in gene expression among the cells that do not reflect the biological differences related to the disease or disorder. A conventional strategy is to regress out the covariate effects for each gene one at a time via a mixed effects model such as MAST [10] and NEBULA [11] where there is a random effect for each individual. However, this regression might be inaccurate since each gene is sparsely sequenced, detrimentally impacting the downstream DE analysis. Hence, an alternative strategy is to use a dimension-reduction method via matrix factorization such as GLM-PCA [12], ZINB-WaVE [13], or scGBM [14]. These methods pool information across genes to remove confounding covariates' effects more effectively. However, a naive application of DE testing on the dimension-reduced scRNA-seq data has been observed to inflate the Type-1 error [15]. This inflation occurs because the dimension reduction introduces correlations among genes that contaminate the signal. Null genes could seem significantly (and erroneously) differentially expressed after a dimension reduction if the test is not performed with care. (See an illustration of this phenomenon in Figure S1.) The main focus of our paper is to solve this statistical dilemma of how to perform differential testing after dimension reduction. We note that another recent line of work has developed differential testing among

cells after a dimension reduction via deep variational autoencoders [16, 17]. The authors overcome the aforementioned pitfall of DE testing after dimension reduction by leveraging the inherent randomness in autoencoders and adding pseudocounts to avoid deeming genes with a small log-fold change as significant. However, we do not pursue this direction since matrix factorizations offer practitioners a more transparent and interpretable framework.

We design the exponential-family SVD differential expression (eSVD-DE) to overcome these two main obstacles, which extends our previous work [18]. Importantly, our method infers the differential expression based on the posterior distribution after performing a dimension reduction, which helps counteract the Type-1 error inflation. This combination of matrix factorization, the posterior distribution, and a test statistic designed to assess differential expression among individuals enables eSVD-DE to better detect DE genes in cohort studies compared to current methods. The eSVD-DE also enables model diagnostics to assess if the assumed statistical model is appropriate for modeling the scRNA-seq dataset.

In this paper, we focus on testing for cells of a particular cell type. We show that eSVD-DE can find reproducible signals in multiple pairs of cohort datasets, either across various cell types between two independent studies of idiopathic pulmonary fibrosis (IPF) in the human lung [6, 19] or within a study of non-inflamed and inflamed cells studying ulcerative colitis in the human colon [20]. We also show that eSVD-DE can find novel DE genes across different cell types in a dataset studying autism [21]. Altogether, these analyses provide evidence that eSVD-DE is a valuable tool for investigating differential expression among cohort studies that will become more prevalent as high-throughput sequencing technologies are applied to large cohorts.

Results

Overview of eSVD-DE for differential expression testing

eSVD-DE performs DE by first projecting the cells onto a low-dimensional manifold while removing the effects of covariates. This step is the cornerstone and namesake of our method. Our dimension reduction follows previous dimension-reduction work such as GLM-PCA [12], ZINB-WaVE [13], and scGBM [14], where we embed the cells via the Poisson distribution based on the gene expression and the covariate variables (Figure 1A). Importantly, these covariates contain an intercept term, the log sequencing depth (computed as the log of the total counts per cell), the case-control indicator, and covariates that could be potential confounders, such as clinical covariates of each individual (sex, age, and smoking status). The eSVD-DE learns a coefficient matrix that removes the effects of the covariates as well as the low-dimensional embedding of "residuals." While this low-dimensional embedding can be helpful for a wide variety of applications downstream, we focus on modifying our previous work [18] to perform DE analysis for cohort data. We advocate using a matrix-factorization-based approach to perform this dimension reduction over deep learning approaches since our approach enables conventional model diagnostics to assess and tune eSVD-DE (Figure 1B).

After fitting a low-dimensional embedding, we cater the following procedure to test for DE among individuals, an aspect absent from our previous work [18]. First,

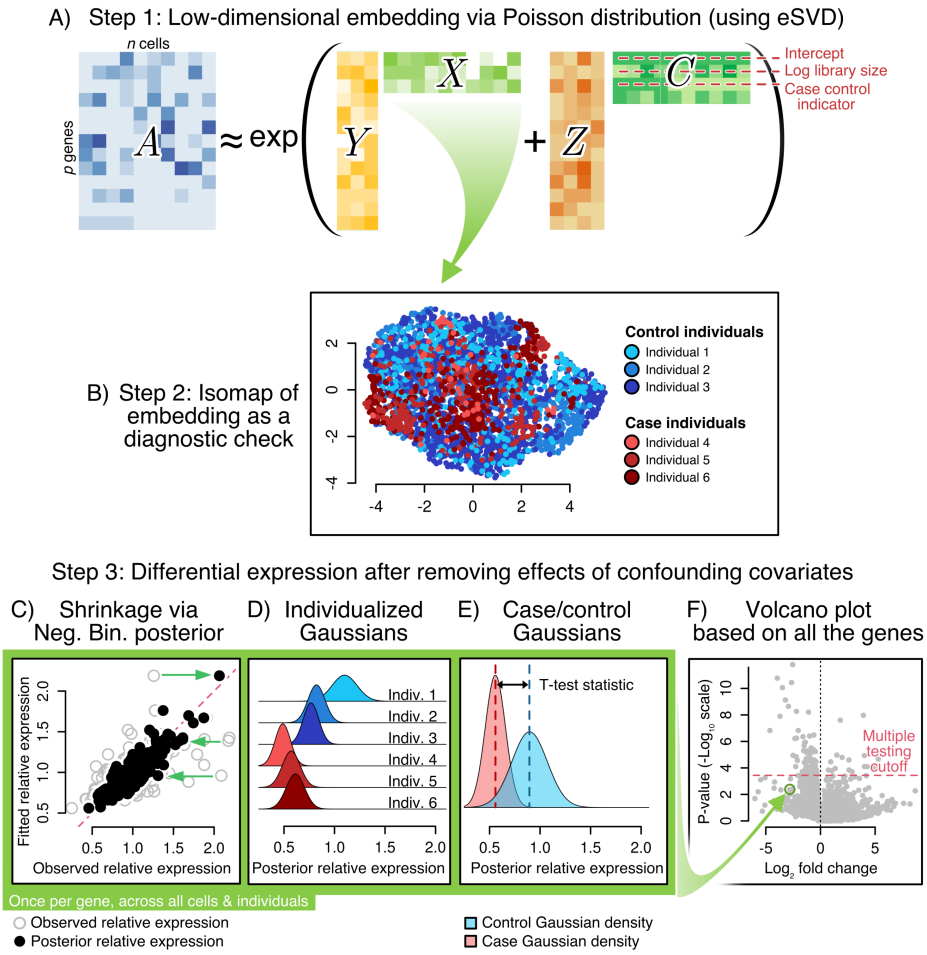

Figure 1 **A** Schematic of the eSVD-DE's matrix factorization, where the observed scRNA-seq data is modeled as a sum of two low-rank matrices, one for the covariates and one for the cells' latent vectors, with an exponential link function (for the Poisson distribution). **B** The cells' latent vectors can be used for diagnostic checks, such as visualization via Isomap. **C** To account for overdispersion and over-fitting of the dimension reduction, shrink each cell via the negative binomial distribution's posterior mean. **D** Represent each individual by a Gaussian distribution among the individual's cells. **E** Compute a test statistic analogous to the T-test after aggregating cells from the cases or control individuals in the cohort. C through E are performed for each gene. **F** Volcano plot, showing a multiple testing cutoff to determine the significant DE genes.

we adjust the predicted relative expression for each cell's gene expression using its posterior according to the Gamma-Poisson distribution (i.e., Negative Binomial, Figure 1C). This posterior is computed by first assessing the overdispersion of each gene, which measures how much extra variability exists between the assumed Poisson fit in the low-dimensional embedding and the observed data, analogous to works like SAVER [22] and totalVI [23]. Importantly, this posterior distribution is computed to be the relative expression after accounting for the contributions of the confounding covariates. Then, for a particular gene, we summarize each individual's cells as a Gaussian distribution following the Central Limit Theorem and compute the T-test statistic reflecting the collective difference between the Gaussians from case individuals to those from control individuals (Figure 1D-E). Finally, we use a multiple-testing procedure based on empirical null distribution to report the DE

genes, which has been successful in other settings to account for possible model misspecification [24] (Figure 1F). More details of the eSVD-DE procedure are described in the Appendix.

eSVD-DE relies on three primary statistical assumptions: (1) scRNA-seq data can be appropriately modeled through the Gamma-Poisson distribution, (2) the effects of confounding covariates can be effectively removed through a GLM framework, and (3) the DE genes show significant differences in means between case and control individuals after accounting for the individual-level covariates. Towards the first assumption, our posterior distribution effectively models counts through a Negative Binomial distribution, which has been justified for modeling scRNA-seq data [25] and has served as the foundation for many methods [13, 22, 26]. Towards the second assumption, many methods from the bulk RNA-seq data such as edgeR [9] and DESeq2 [8] have found tremendous success regressing out covariates through the GLM framework. Towards the last assumption, we note that there are existing methods such as IDEAS [27], scDD [28], and waddR [29] that test for differential *distributions* instead of differential mean expressions. However, we have found it more challenging to generalize these results when comparing different datasets of similar biological systems, as differentially distributed genes are not neatly characterized by over-/under-expression.

Design of simulation studies

The main message we wish to convey in this simulation section is two-fold: 1) testing for differential expression among individuals is fundamentally different from among cells, and 2) eSVD-DE's framework enables more accurate inference due to its usage of dimension reduction and posterior correction. Figure 2A conceptualizes the first point. Large within-individual variability hinders "pseudobulk" DE methods that sum the gene expressions across all the cells originating from the same individuals since these methods do not account for the variability of expression within an individual. On the other hand, large between-individual variability hinders most existing methods designed for testing DE genes among cells from scRNA-seq data. This is because even if the difference in mean expression is insignificant on the cohort level, many cells from a small subset of individuals can yield highly significant differences in mean expression on the cell level. Since previous work has explained the importance of accounting for within-individual variability [27, 31, 30], we focus on a simulation that illustrates the importance of accounting for the between-individual variability here.

Our simulation study contains 700 genes and 20 individuals equally split among cases and controls, where each individual contributes 250 cells. The cells have gene expression profiles that are impacted by covariates correlated with the case-control status (such as age, sex, and tobacco use) and are sampled from a Gamma-Poisson distribution (Figure 2B). Overall, the true generative model is more complex than the statistical model assumed by eSVD-DE to not give eSVD-DE an unfair advantage in our benchmarking experiment. Nonetheless, the data is generated such that 50 genes are drastically more differentially expressed among cases and controls on the cohort level than the remaining 650 genes, barring confounding effects. This true signal can be visualized via an Isomap [32, 33], where there is no apparent

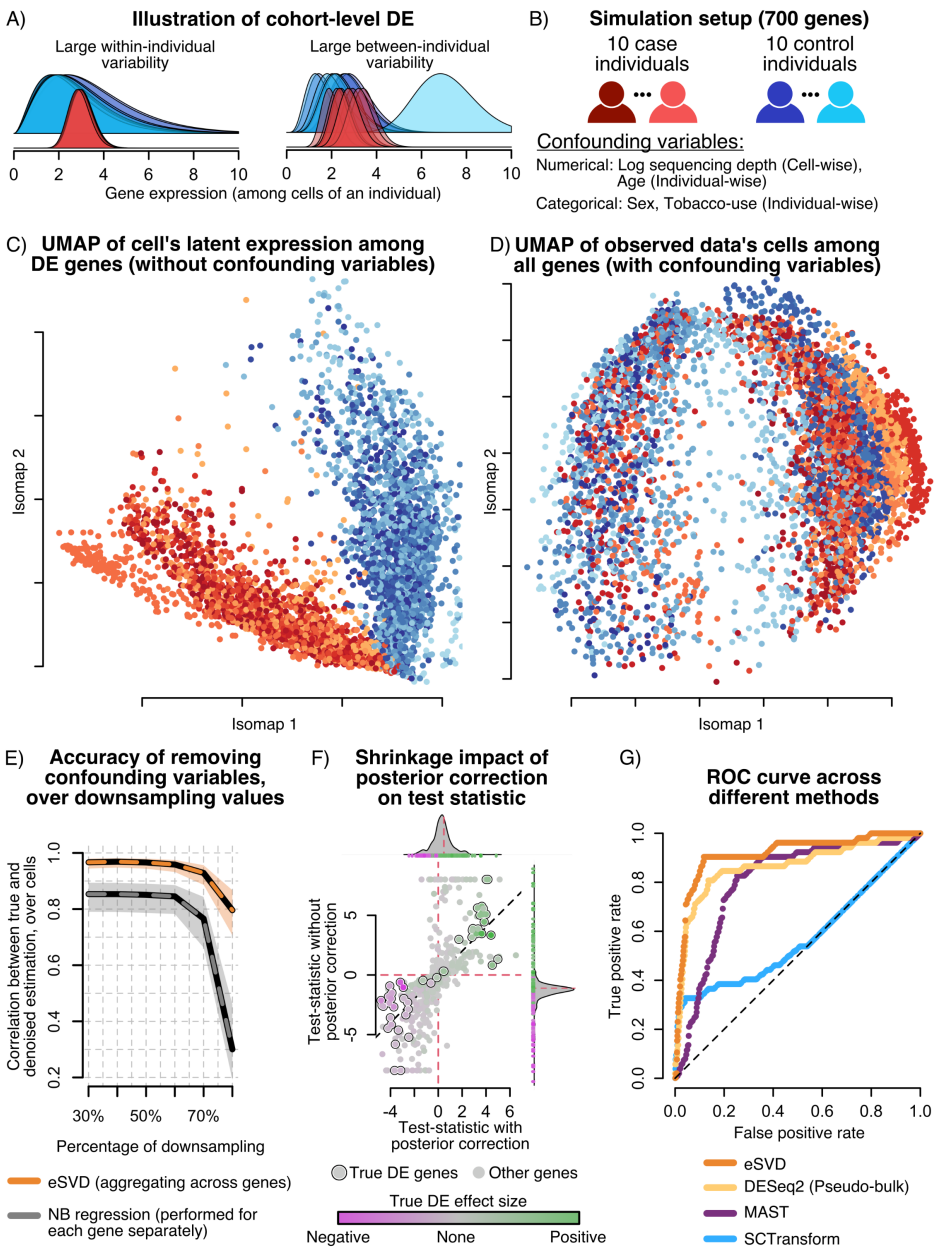

Figure 2 **A** Illustration of challenges for cohort-wide DE testing. **B** Setup for our simulation setup. **C** Isomap of the cells based on the true DE genes' expression before introducing the confounding variables. No individuals concentrate tightly in any region on the Isomap manifold, and there is a strong separation between the cases (shades of red) and controls (shades of blue). **D** Isomap of the observed data based on all the genes. Cells from the same individual concentrate in the embedding, suggesting that confounding covariates additionally drive the difference in expression profiles among individuals. **E** Downsampling experiment, demonstrating that by pooling information across genes, eSVD-DE outperforms gene-by-gene Negative Binomial regression for regressing out covariate effects. **F** Illustration of the importance of shrinkage, where the x-axis and y-axis represent each gene's test statistic with and without posterior correction, respectively. The genes are colored by their true log-fold change, of which the circled genes denote the top 50 genes with the highest true log-fold change. **G** ROC curve comparing four different methods, illustrating that eSVD-DE has more power than competing methods.

stratification within the case or control individuals (Figure 2C). However, when the

covariate effects of age, sex, and tobacco use are also included, there is obvious confounding of which genes are differentially expressed (Figure 2D). Specifically, the cells from the same individual concentrate in different regions of the embedding. We choose to use the Isomap here, as opposed to the more commonly used UMAP [34], since the Isomap has well-studied statistical properties [35, 36, 37]. This quality is valuable when we use these visualizations to diagnose our hypothesis testing framework instead of an exploratory tool, as we will illustrate later.

Simulations verify that eSVD-DE effectively remove covariate effects in the presence of sparsity

We first show that by pooling information across genes via the dimension-reduction framework, eSVD-DE removes the confounding variables' effects more accurately compared to a Negative Binomial (NB) regression applied separately for each gene (Figure 2E). To demonstrate this, we incrementally downsample the data to reduce the signal size of the DE genes. As the amount of downsampling increases, it will be more difficult for a method to remove the confounding effects properly. We measure the accuracy of how well the confounding effects were removed by computing the Pearson correlation between each cell's vector of true natural parameters across the 700 genes and its vector of estimated natural parameters after the confounding effects have been removed. We observe that at starting at 30% downsampling, eSVD-DE removes the confounding effects more accurately than the gene-by-gene NB regression (0.95 to 0.87). Additionally, although both methods drop in accuracy for downsampling levels of 60% or more, eSVD-DE is still more accurate than the gene-by-gene NB regression. This finding is statistically intuitive since when genes are sparsely observed, there is not enough information within a specific gene to accurately estimate the NB regression coefficients. However, by pooling information across genes, the coefficients for the confounding variables are better estimated.

Simulations verify that eSVD-DE's posterior enables gene-specific discoveries

We next illustrate that the posterior correction performed by eSVD-DE is important for DE testing on the cohort level (Figure 2F). We plot the test statistic derived directly from the low-dimensional embedding against the test statistic derived from the posterior-corrected relative expression, where we mark the 50 genes the largest true log-fold change. In particular, the former represents the prototypical analysis of applying DE after denoising the data by a dimension-reduction method, and we observe that many null genes display large test statistics (x-axis), confirming the findings of previous work that observes an inflation of Type-1 error for such procedures [15, 31, 38]. However, by adjusting the relative expressions using the posterior mean, many null genes have drastically smaller test statistics (y-axis). This phenomenon occurs because the assumed linear model between the covariates and the gene's natural parameter does not sufficiently capture more complex non-parametric relationships often displayed among individuals, distorting the lower-dimensional embedding of the "residuals." This distortion has an adverse effect when genes vary substantially in sparsity – sparsely observed genes are denoised by projecting the cells onto the incorrect manifold.

Simulations verify that eSVD-DE's low-dimensional embedding improves power over other methods

Lastly, we illustrate the eSVD-DE has more power than other conventional methods to test for DE in cohort-wide scRNA-seq data through our simulation. Since different methods estimate sets of DE genes with dramatically different sizes for a particular FDR cutoff, we plot the entire curve of true positive rate (TPR) and false positive rate (FPR) over all possible cutoffs (Figure 2G). Here, we compare against three other methods: DESeq2 [8] (i.e., the prototypical method representing "pseudobulk" analyses), MAST [10] (i.e., the commonly used method using mixed-effect models where there is a random effect among individuals) and SCTransform [39] (i.e., the prototypical method of performing DE ignoring the individual structure). We observe that eSVD-DE has the highest power compared to the three other methods. DESeq2 performs the best among the three competing methods since the averaging among cells of an individual dramatically reduces the estimation variability, while MAST performs the next best since it accounts for the individual structure.

We perform a separate simulation in the Appendix (Figure S2 and S3) to demonstrate that eSVD-DE does not inflate the Type-1 error among true null genes.

eSVD-DE enables diagnostics to assess the performance of removing covariate effects

Moving beyond simulations, we now investigate how well eSVD-DE removes confounding effects in scRNA-seq datasets and how the dimension-reduction framework enables conventional model diagnostics. To demonstrate this, we investigate a broad collection of scRNA-seq datasets from diverse tissues but focus here on the 8909 T-cells from a dataset of lung cells containing 10 healthy individuals and 24 individuals with IPF sequenced using the 10x Chromium single-cell platform, henceforth called the "Adams" dataset [6]. We focus on 6969 genes for our analysis, including 5000 highly-variable genes, the genes reported to be DE by the authors as well as the 1003 human housekeeping genes [40]. When visualizing these cells via Isomap, we can see a clear separation of the case and control individuals, suggesting there are many genes with separable expression patterns (Figure 3A). However, we also see that individual-level covariates like sex and smoking status also are locally concentrated in different regions of the Isomap. We do not wish to report genes as differentially expressed if the differences are induced only by sex differences or smoking.

After applying eSVD-DE, we see that the resulting Isomap of the fitted embedding no longer carries expression patterns correlated with the individuals, sex, or smoking status (Figure 3B). If biologists prefer quantitative diagnostics, the practitioner can purposely omit a small percentage of values from the scRNA-seq count matrix before applying eSVD-DE and assess how correlated the predicted values are to these omitted values. This strategy was used successfully in our previous work [18]. For all these reasons, we advocate the matrix-factorization strategy in the eSVD-DE as opposed to deep-learning alternatives where there are fewer interpretable diagnostics available. The remaining diagnostics and the Isomaps for other scRNA-seq datasets are shown in the Appendix (Figure S4).

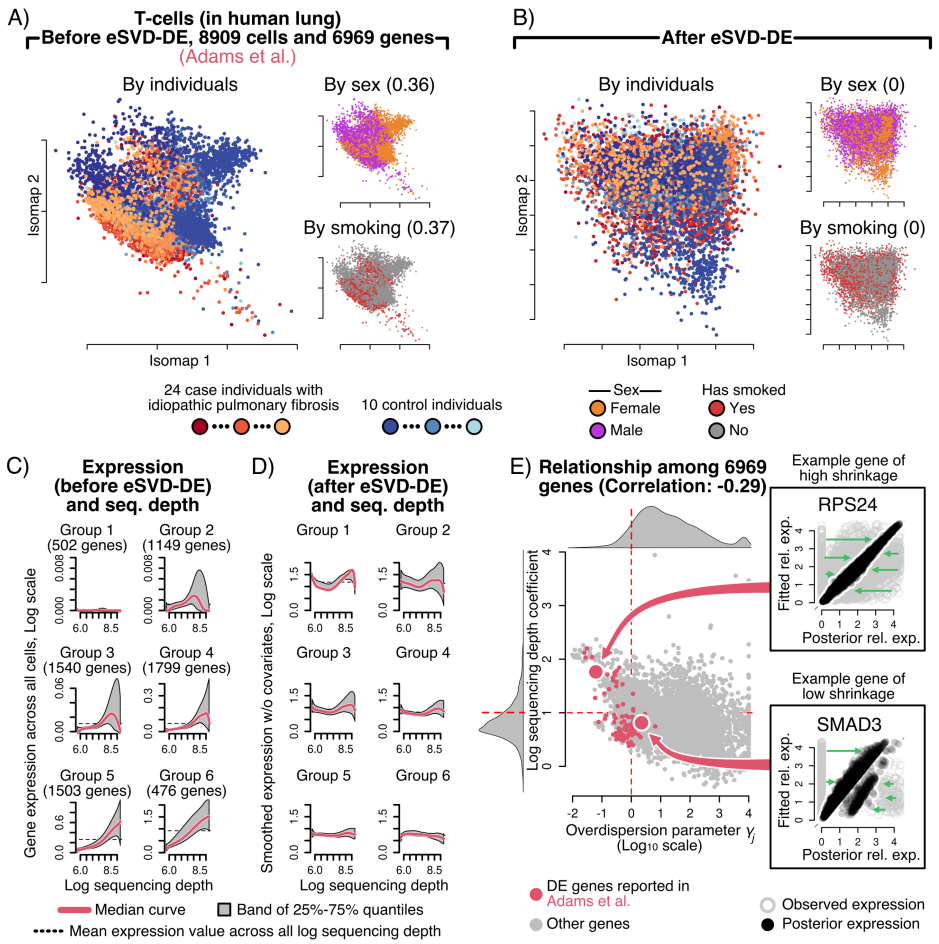

Figure 3 **A** Isomap of the leading principal components for the T-cells in the Adams dataset, shown by individuals, sex, and smoking status. The number in the insets denotes how correlated the individual-level covariate is correlated with the leading principal components. **B** Isomap of the cell embedding after applying eSVD-DE, demonstrating that confounding covariate and individual effects have been removed. The number in the insets denotes how correlated the individual-level covariate is correlated with the estimated eSVD-DE cell embedding. **C,D** Relationship between gene expression and cells' log sequencing depth before or after eSVD-DE (shown on a log scale), respectively, for groups of 6 genes partitioned by the gene's mean expression relationship after eSVD-DE. After eSVD-DE, each gene becomes more uncorrelated with the cell's log sequencing depth. **E** Relationship between each cell's log sequencing depth and overdispersion parameter, highlighting two genes displaying different shrinkage levels via the posterior mean.

eSVD-DE appropriately adjusts for the sequencing depth using a dimension-reduction approach

Many authors have concluded that appropriately adjusting for the sequencing depth of each cell is one of the most influential aspects of an effective DE method [41, 42]. This adjustment is critical for scRNA-seq data since we do not wish to deem genes as DE simply because specific cells have a larger sequencing depth than others. Instead, genes should be deemed as DE if the case and controls have significantly different expression relative to the cells' sequencing depth. Accounting for this sequencing depth has been the source of numerous debates on how to normalize the cells' expression best [43]. The most commonly used approach is to log-normalize the

gene expressions, but many existing work has cited that this normalization distorts qualitative aspects of the scRNA-seq dataset [12, 39].

In order to quantify how well the sequencing depth effect was removed from the scRNA-seq data, we perform diagnostics analogous to those in [39]. Specifically, we group the genes into 6 bins based on their mean gene expression. We observe a significant relationship between each gene's expression and the cells' sequencing depth before normalizing the scRNA-seq data (Figure 3C). This means without any accounting of the sequencing depth, genes could be deemed significantly differentially expressed solely due to cells having different sequencing depths. However, after fitting the eSVD-DE, this relationship is mainly removed across all bins, discounting the genes with the smallest mean expressions (Figure 3D). Observe that eSVD-DE bears an advantage over other normalization methods such as SCTransform [39] and Scran [44] since eSVD-DE assesses the appropriate sequencing-depth normalization by simultaneously accounting for the cells' gene expression and covariate information through its dimension-reduction framework, an important quality when dealing with cohort data. See Figure S5 for the application of this diagnostic on other datasets, and Figure S6 for how this diagnostic performs when only log-normalization is performed instead of the eSVD-DE in the Appendix.

Lastly, recall that our eSVD-DE framework adjusts the fitted gene expression values by the posterior Negative Binomial distribution. This adjustment relies on first quantifying the overdispersion of each gene, i.e., a higher overdispersion means the gene's expression patterns conform less to the estimated low-dimensional embedding. We hypothesize a negative correlation between the sequencing-depth normalization and the amount of overdispersion. This hypothesis is in line with previous works that investigate this relationship [39], citing that a smaller overdispersion implies that the fitted values are a good approximation of the gene expressions, which often occurs for densely observed genes. This can be visualized as a scatterplot (Figure 3E). Additionally, this plot enables practitioners to survey the amount of shrinkage across all the genes broadly. We highlight RPS24 and SMAD3, 2 genes implicated in previous studies [45, 46, 47], as example genes that are highly and lowly shrunk via the posterior distribution since these genes were . All these aspects collectively support the claim that eSVD-DE is appropriately adjusting each gene's expression by the sequencing depth, which enables us to investigate the performance of the DE analyses downstream.

eSVD-DE recovers reproducible differences between case and control expression across multiple datasets

While it is difficult to assess the validity of DE genes in cohort-wide scRNA-seq data due to the lack of negative control genes, we hypothesize that a reliable proxy to assess the quality of the DE method is to collect two different cohort-wide scRNA-seq datasets of the same system and diseases/disorder and see if genes reported to be significant in one dataset display similar significances in the other. Towards this end, we investigate another dataset of 5286 T-cells of lung cells from 4 healthy individuals and 6 individuals with IPF sequenced using the 10x Chromium single-cell platform of the same 6969 genes here, henceforth called the "Habermann" dataset (Figure 4A) [19]. Since we are primarily interested in comparing eSVD-DE to other DE

methods using this pair of datasets, we do not deploy a multiple-testing procedure to select DE genes but investigate the sets genes with the largest test statistic magnitudes of the same cardinality as those reported by the authors for meaningful comparisons. For starters, the volcano plot on the Adams dataset shows that 20 of the 84 genes with the largest test statistics derived from our eSVD-DE procedure intersect with the 84 DE genes reported by the authors, resulting in Fisher's exact test p-value of 2.2×10^{-21} (Figure 4B). Similarly, the volcano plot on the Habermann dataset shows that 30 of the 157 genes with the largest test statistics derived from our eSVD-DE procedure intersect with the 157 DE genes reported by the authors, resulting in Fisher's exact test p-value of 3.1×10^{-20} (Figure 4C). Additionally, since housekeeping genes constitute genes primarily responsible for basic cellular functions and are stably expressed regardless of cellular condition, we hypothesize that these genes should not carry substantial differential expression patterns [40]. This phenomenon is demonstrated through the volcano plots (Figure 4B,C).

When comparing different DE methods, we ask if the leading genes derived from the Adams dataset using eSVD-DE are either (1) genes reported by authors of the Habermann dataset or (2) themselves the leading genes derived from the Habermann dataset using eSVD-DE, or vice-versa. If the size of this intersection is larger than those derived by other DE methods in either scenario, we would have partial evidence that eSVD-DE is more consistently recovering reproducible signals across both datasets. We compared against DESeq2 whereby cells of each individual are aggregated into "pseudobulk" expressions prior to performing DESeq2 [8] for this investigation. We highlight this pseudobulk procedure specifically since this procedure was reported to be more reliable for scRNA-seq with biologically-replicate samples, but its reliability for cohort data was not investigated [4]. For this comparison, we select the 226 genes with the largest test statistics in magnitude from either the Adams or Habermann dataset using either eSVD-DE or DESeq2 since the 84 reported DE genes for the Adams dataset and 157 reported DE genes for the Habermann dataset yield 226 unique genes. The upset plot shows that eSVD-DE yields larger intersections between the two datasets than DESeq2 (Figure 4D).

Additionally, we hypothesize that if a DE method recovers reproducible signals, the DE genes test statistics should be positively correlated between the two datasets. Indeed, if a reported DE gene has a positive log-fold change in one dataset, the other dataset should also show evidence of a positive change. On the other hand, we hypothesize that genes unrelated to the disease/disorder should be uncorrelated test statistics between the two datasets. This hypothesis would be biologically justifiable, as the log-fold change of a gene unrelated to disease/disorder would be determined by random chance. We observe these relationships for eSVD-DE's test statistics (Figure 4E). In contrast, the correlation for DESeq2's test statistic among the reported DE genes is near-zero (Figure 4F). We suspect all the above phenomenons are likely driven by pseudobulk methods' lack of accounting for the within-individual variability.

To further support our empirical claims regarding eSVD-DE, we also study cells from 18 individuals with ulcerative colitis (UC) and 12 healthy individuals across multiple cell types in the colon and 5713 genes, henceforth called the "Smillie" dataset [20]. Each individual with UC contributed cells from inflamed and non-inflamed colon biopsies, and each healthy individual contributed two biologically

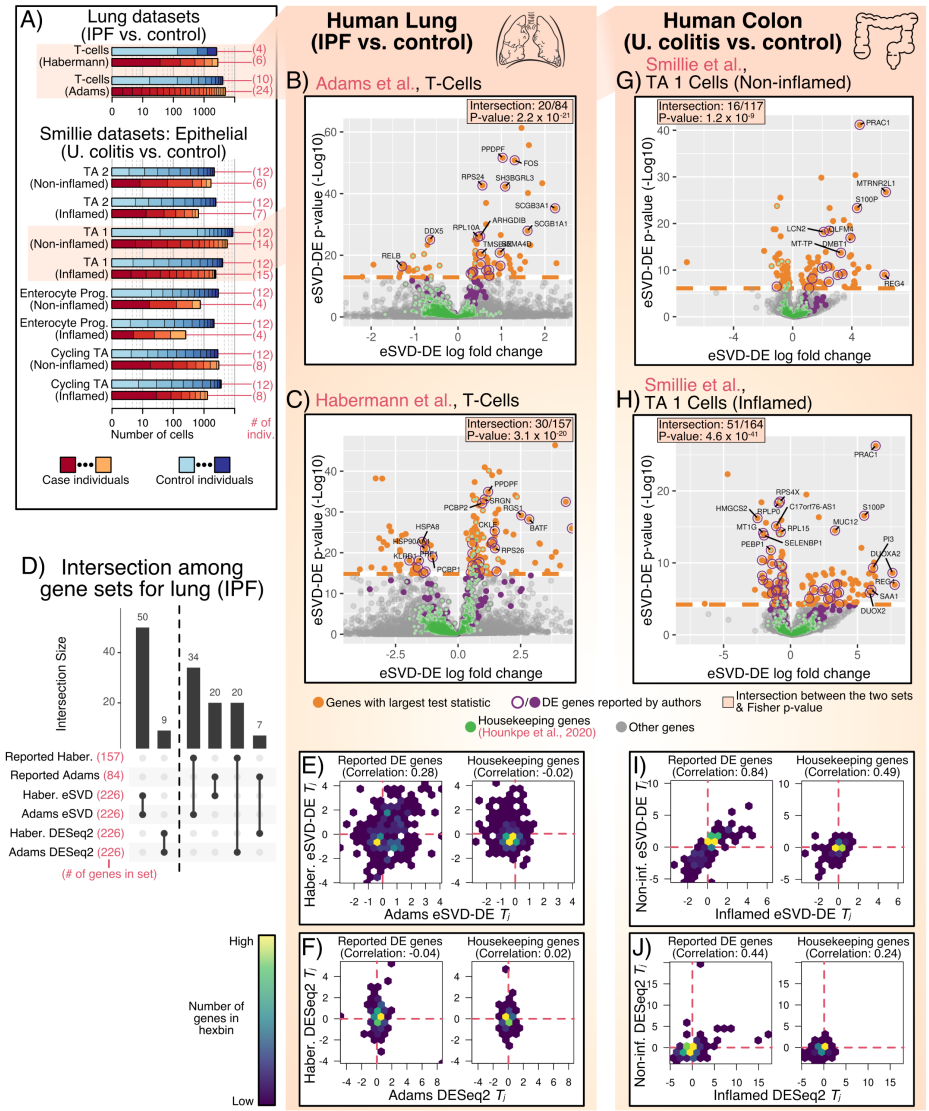

Figure 4 **A** Number of case and control individuals and the total number of cells across cell types and datasets. The x-axis denotes the total number of cells on a log scale, while the partitioning of case and control cells denotes the number of cells from each individual relative to the total number of cells. **B,C,G,H** Volcano plot, where the set of genes with the largest test statistics (having the same size as the original author's set) is depicted in orange. The genes reported by the respective authors or housekeeping genes are in purple and green, respectively. The Fisher exact test's p-value between the enrichment of eSVD-DE's DE genes and the author's reported DE genes is also reported. **D** Upset plot showing the intersection between pairs of DE genes, either the reported DE genes in the Habermann and Adams dataset and the estimated DE genes using eSVD-DE or DESeq2. **E,I** Hexplots showing the correlation between genes' eSVD-DE test statistics across datasets, either the originally reported DE genes in either dataset or the housekeeping genes. **F,J** Similar to **E,I** but showing the gene's DESeq2 test statistics.

replicated samples from biopsies in analogous colon regions. In our analysis, we treat each tissue sample as a different "individual" and perform two DE analyses – one of non-inflamed samples from individuals with UC against healthy samples (i.e., the "non-inflamed" analysis) and another of inflamed samples from individuals with UC against healthy samples (i.e., the "inflamed" analysis). Importantly, we split the

healthy tissue samples so each healthy sample is only involved in one of the two DE analyses. While we apply this pair of analyses on multiple cell types in this biological system, we focus on the analysis of transit amplifying 1 (TA 1) cells here (Figure 4A). Similar to before, we see that the volcano plots for both the non-inflamed and inflamed DE analysis yield highly-enriched intersections between the genes with the largest test statistics and the DE genes reported by the authors, as well as near-zero enrichment of the housekeeping genes (Figure 4G,H).

The authors of the Smillie dataset observed a high correlation among the test statistics of reported DE genes between the non-inflamed and inflamed DE analyses when aggregating across all the cell types, suggesting that the transcriptomic signature of UC precedes inflammation [20]. We hypothesize that a higher-powered DE analysis of cohort-wide scRNA data should reveal a strong correlation even when focusing on only one cell type. Indeed, we see this positive correlation among eSVD-DE's test statistics ($\rho = 0.84$, Figure 4I). Additionally, while we see a positive correlation among the housekeeping genes ($\rho = 0.49$), many of such genes have a near-zero test statistic in the non-inflamed analysis. This observation suggests that the differences in housekeeping genes' expression are induced more by the inflammation of the tissue rather than a biological mechanism disrupted by ulcerative colitis. In contrast, when we apply a similar pair of analyses using DESeq2 on pseudobulk data, the correlation among the reported DE genes is substantially lower (Figure 4J).

eSVD-DE detects DE genes highly enriched with previously-annotated genes

Having demonstrated all the advantages of eSVD-DE, we hypothesize that our method leads to novel cell-type specific DE discovery. Towards this end, we analyze brain single-cells from controls and case individuals who had been diagnosed with autism spectrum disorder (ASD), partitioned among many cell-types [21], henceforth called the "Velmeshev" dataset (Figure 5A). We focus on this particular system since the Simons Foundation Autism Research Initiative (SFARI) routinely curates a list of autism risk genes based on genetic studies from many publications [48], which can provide relevant genes to compare our eSVD-DE results against. Additionally, a recent bulk-DE analysis of ASD provides a high-quality and independent list of genes to compare against, which we will call the "bulk DE genes" henceforth [49]. A priori, we would not expect any method deployed on the Velmeshev dataset to fully match these lists because our analysis is cell-type specific and the SFARI list are genetic risk genes, which do not always lead to differential expression.

When analyzing the 12,984 cells in layer 2/3, we observe that eSVD-DE estimates 331 DE genes (among the 7055 genes used in the analysis) using an empirical FDR cutoff of 0.05, where we calibrate the p-value according to an empirical null distribution to account for potential misspecification [50] (Figure 5B). Here, 42 and 95 of these genes are among the SFARI and the bulk DE genes, respectively (among the total 800 and 1556 genes, respectively). Comparing the bulk DE genes to our eSVD-DE genes (which are both derived from transcriptomics data), we obtain a significant Fisher p-value of 0.002. Additionally, when we compare DESeq2 for the same cells to find DE genes (144), only 19 and 30 genes overlapped with the SFARI

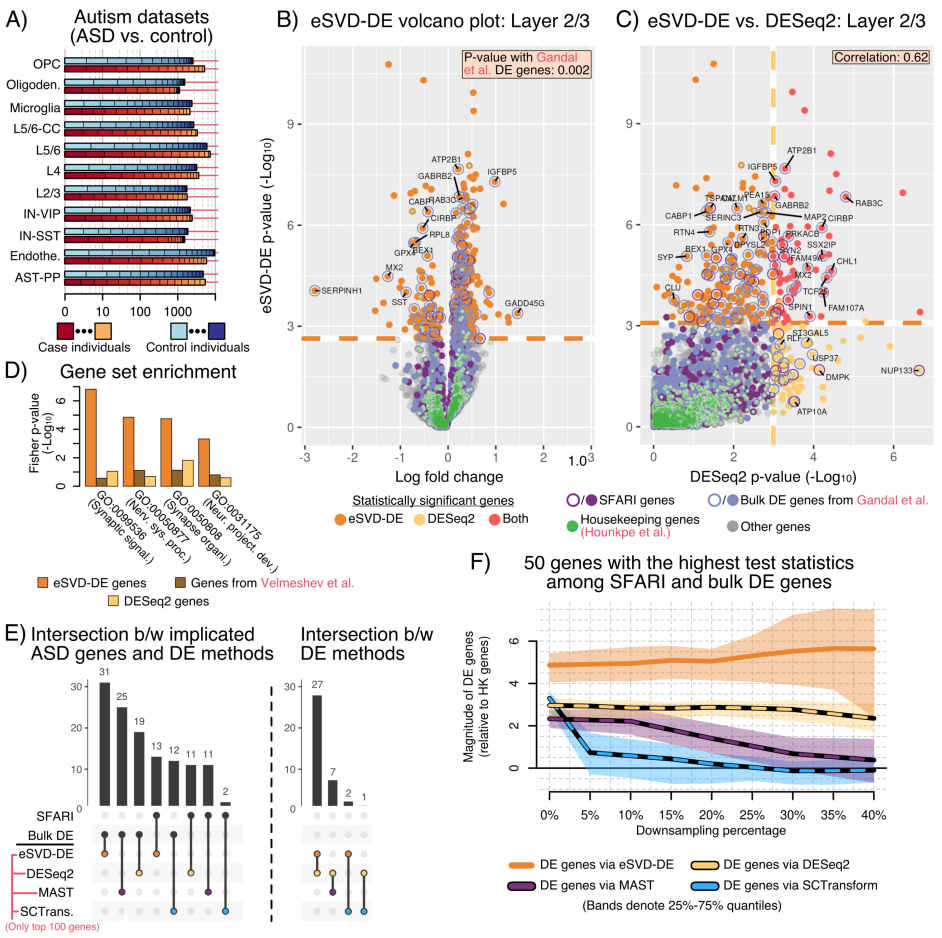

Figure 5 **A** Number of case and control individuals and the total number of cells across cell types for the Velmeshev dataset, analogous to Figure 4**A**. **B** Volcano plot showing eSVD-DE's results for layer 2/3 cells, where the set of genes exceeding the empirical FDR cutoff are depicted in orange. The SFARI, bulk DE, and housekeeping genes are in purple, blue, and green, respectively. The Fisher exact p-value between the selected genes via eSVD-DE and the bulk DE genes is noted. **C** Comparison of the eSVD-DE p-values against the DESeq2 p-values, and the genes exceeding an FDR cutoff for DESeq2 are shown in yellow. The correlation between the two sets of negative $\log_{10}$ p-values is noted. **D** Increase in enrichment for relevant GO terms among the 331 DE genes found by eSVD-DE, the 144 DE genes found by DESeq2, and the 109 genes initially reported by the authors. **E** Upset plot comparing the SFARI or bulk DE genes with the top 100 DE genes estimated by eSVD-DE, DESeq2, MAST, or SCTransform. **F** Downsampling experiment, illustrating the stability of the 50 genes with the largest test statistics in magnitude among the SFARI or the bulk DE genes as the dataset is artificially downsampled. The values on the y-axis is the ratio of the test statistic between these genes and the housekeeping genes.

and the bulk DE genes, respectively (Figure 5C). This finding demonstrates that eSVD-DE can find more DE genes, which are more relevant, compared to DESeq2. We report analogous plots in the Appendix when comparing against other methods, such as MAST and SCTransform (Figure S7). We note that we do not compare against the DE genes found by the authors of the data themselves here, as they had used MAST to find their DE genes. As our simulations beforehand demonstrated, MAST is not necessarily a reliable "gold standard" to compare against. However, the 331 DE genes found by eSVD-DE are also much more enriched in GO terms that are plausibly related to ASD when compared to the DESeq2 genes or the

original genes reported for layer 2/3 (Figure 5D). All these observations suggest that eSVD-DE is well-equipped to uncover meaningful results.

Next, we investigated qualitative differences among the methods compared to the SFARI and bulk DE genes. We enable fair comparisons among the different methods by finding the top 100 genes with the smallest p-values for each method. We then counted how many of these 100 genes intersected with the SFARI or bulk DE genes (Figure 5E). eSVD-DE has the highest overlap with both sets (31 with the bulk DE, 13 with the SFARI genes), followed closely by DESeq2 and then finally by MAST and SCTransform. In fact, eSVD-DE and DESeq2 have an overlap of 27 genes. The observations from Figure 5D and E combined suggest that eSVD-DE inherits many advantages of pseudobulk methods while improving these methods by accounting for within-individual variability. In the second investigation, we asked how stable each method was as the dataset was gradually downsampled. Indeed, as the scRNA-seq dataset becomes sparser, the relative difference between "strongly-expressed" DE genes and null genes becomes fainter, and we would expect any method to degrade in performance. To investigate this, we take the 50 genes with the largest test statistics in magnitude among the SFARI and bulk DE genes for each method and ask how their mean test statistic (relative to the housekeeping gene's test statistic) diminishes as the scRNA-seq data is downsampled. We use the housekeeping genes as a proxy for the null genes, which is a reasonable choice given Figure 5B. As the downsampling percentage increases (i.e., the signal becomes fainter), MAST and SCTransform lose their ability to distinguish between formally highly-differentiable genes and housekeeping genes. This finding makes sense, as MAST and SCTransform estimate the DE genes based on a gene-by-gene regression, meaning no information is shared between genes. On the other hand, DESeq2 and eSVD-DE are surprisingly stable – they can demonstrate a clear separation between the 50 DE genes and the housekeeping genes even when the data is downsampled at 40%. This finding is also sensible, as DESeq2 aggregates cells among individuals, and eSVD-DE pools information between genes; both strategies yield robustness against sparsity. However, eSVD-DE is still preferred over DESeq2 here, as the separation between the DE genes and housekeeping genes is much higher for eSVD-DE.

We include additional diagnostic plots via Isomaps, volcano plots of the DE genes, and the GO analysis results in Appendix (Figure S5, and S8 through S10).

Conclusions

We demonstrate the nuances of performing cohort-wide differential testing for single-cell RNA-seq data. The difficulty primarily stems from individual-level confounding covariates, which can be difficult to remove using typical DE strategies of regressing out their effects gene-by-gene. Instead, eSVD-DE pools the information across genes to remove the confounding covariates, yielding empirical performance that is more promising than current pseudobulk DE methods or DE methods currently used for scRNA-seq data when no prevalent individual-level covariates are present. We achieve this through a matrix factorization strategy, estimating the coefficients associated with the covariates and each cell's and gene's latent vectors. We then shrink the estimated denoised gene expressions via the posterior mean according to the Gamma-Poisson distribution. This strategy helps dampen potential

over-smoothing effects induced by the matrix factorization. We then deploy a test statistic designed to test for DE genes on the individual level instead of the cellular level. Note that our procedure can be used concurrently with IDEAS [27], which aims to find differential *distributions* between cases and controls, which could be more challenging to interpret than differential means. However, we do not pursue this direction in this paper. We hope that eSVD-DE would be beneficial for inspiring cohort-wide DE tests for future single-cell assays beyond scRNA-seq and single-cell eQTL analyses where abundant individual-level covariate effects must be adequately removed. Additionally, we are curious about broader settings where instead of having case and control individuals within a cohort, we are interested in testing if continuous covariates such as age have a substantial transcriptomic impact on specific cell types, as discussed in [51].

Appendix

Overview of eSVD-DE model

Statistical model

Let $A \in \{0, 1, \ldots, \}^{p \times n}$ denote the observed count matrix with n cells and p genes, and $C \in \mathbb{R}^{n \times r}$ denote the observed r covariates for the n cells. Importantly, certain columns of C would be the following:

- **Intercept**: Let $C_{\cdot,1} = 1$. We'll call this column $C_{\cdot,(\text{int})}$.
- **Log sequencing depth**: Let $s_i = \sum_{j=1}^{p} A_{ji}$. Then, let $C_{i,2} = \log(s_i)$ for all $i \in \{1, \ldots, n\}$. We'll call this column $C_{\cdot,(\text{lib})}$.
- **Case-control status**: Let $C_{i,3} \in \{0, 1\}$ depending on whether or not cell $i \in \{1, \ldots, n\}$ is associated with a case or control individual. We'll call this column $C_{\cdot,(\text{cc})}$.
- **Others**: The remaining columns of C could contain numerical covariates associated with the cells: which region of the body the cell was sampled from, the age of the corresponding individual, etc. We assume the categorical covariates have already been transformed via one-hot encodings (i.e., categorical variables with k levels are transformed using $k-1$ indicator variables). In contrast, assume the numerical covariates are standardized to have a standard deviation of 1. We have found it beneficial to not center the numerical covariates around 0. This way, the resulting estimated coefficients in Z are easier to interpret.

Optionally, practitioners may include one-hot encoding vectors for which cells originate from which individuals. In our paper, we have found this to be optional and sometimes detrimental to the fit. This is because the inclusion of such one-hot encoding vectors result in a collinear matrix C. This preparation of the covariate matrix C is primarily handled by the function `eSVD2::.reparameterization_esvd_covariates` in our codebase.

The statistical foundation of eSVD-DE is the following hierarchical model where each entry of A_{ji} is modeled as,

$$\left(A_{ji} | \lambda_{ji}\right) \sim \text{Poisson}(\ell_{ji} \cdot \lambda_{ji}), \quad \text{and} \quad \lambda_{ji} \sim \text{Gamma}(\alpha = \mu_{ji}/\gamma_j \; ; \; \beta = 1/\gamma_j), \quad (1)$$

(where α and β denote the shape and rate parameters of a Gamma distribution respectively) for gene $j \in \{1, \ldots, p\}$ and cell $i \in \{1, \ldots, n\}$, where the low-dimensional mean matrix is $\mu \in \mathbb{R}^{p \times n}$ where

$$\mu_{ji} = \exp\big((Y_{j,\cdot})^\top (X_{i,\cdot}) + Z_{j,(\mathrm{cc})} \cdot C_{i,(\mathrm{cc})}\big), \tag{2}$$

for $X \in \mathbb{R}^{n \times k}$ and $Y \in \mathbb{R}^{p \times k}$, and ℓ_{ji} denotes the covariate-adjusted sequencing depth,

$$\ell_{ji} = \exp\big((Z_{j,\text{-}(\mathrm{cc})})^\top C_{i,\text{-}(\mathrm{cc})}\big), \tag{3}$$

and $\gamma_j > 0$ denotes the overdispersion parameter for gene j which controls the variability of gene j. Here, the subset notation "$-$(cc)" means that we exclude the "(cc)" column from the corresponding matrix. In the terminology of previous work [22, 25], we interpret μ_{ji} as the "predictable" gene expression (i.e., expression of gene j in cell i that can be predicted from other cells and genes, which we interpret as a low-dimensional manifold), and λ_{ji} as the biological relative expression of gene j in cell i. We observe the count matrix A as well as the covariates C, and we need to estimate the cells' latent embedding X, the genes' latent embedding Y, the covariates' coefficients Z as well as the overdispersion vector γ.

Regarding the Gamma distribution. Our parameterization of the Gamma distribution is inspired by the constant Fano factor model used in SAVER [22]. Specifically, for the Gamma distribution in (1),

$$\mathbb{E}[\lambda_{ji}] = \mu_{ji}, \quad \text{and} \quad \mathbb{V}[\lambda_{ji}] = \gamma_j \cdot \mu_{ji},$$

meaning the variance scales proportionally with the mean. Additionally, it can be derived that the marginal distribution of A_{ji} is a negative-binomial,

$$A_{ji} \sim \mathrm{NB}\big(r = \mu_{ji}/\gamma_j \ ; \ p = \ell_{ji}/(\ell_{ji} + 1/\gamma_j)\big),$$

meaning the moments of A_{ji} marginally are

$$\mathbb{E}[A_{ji}] = \ell_{ji} \cdot \mu_{ji}, \quad \text{and} \quad \mathbb{V}[A_{ji}] = (\ell_{ji}\gamma_j + 1) \cdot \ell_{ji} \cdot \mu_{ji}.$$

These equations demonstrate that γ_j measures overdispersion – the larger γ_j is, the larger the variance of A_{ji} is.

Additionally, we can derive that the posterior distribution of $\lambda_{ji}|A_{ji}$ is

$$\lambda_{ji}|A_{ji} \sim \mathrm{Gamma}\big(\alpha = A_{ji} + \mu_{ji}/\gamma_j \ ; \ \beta = \ell_{ji} + 1/\gamma_j\big), \tag{4}$$

which will be useful when deriving our method below.

Rationale and justification of statistical model

Our hierarchical model (1) draws inspiration from two literatures. The first literature is the single-cell modeling literature. Various work have shown that hierarchical model (1) that model observed counts as a Poisson distribution originating from a Gamma prior is a statistically sound and empirically justified model of scRNA-seq data [22, 25]. These models are appealing as they explicitly separate the technical noise (i.e., noise incurred by the nature of sequencing, modeled via the Poisson distribution) from the biological noise (i.e., noise naturally occurring among cells, modeled via the Gamma distribution). Other work has provided additional biological explanations on why biological noise is commonplace in organisms [52].

The second literature is the differential-expression literature, which canonically focuses on testing for the differential expression of a gene, one gene at a time. For gene j, methods like DESeq2 [8], MAST [10], and SCTransform [39] regress the observed expression $A_{j1}, \ldots, A_{jn}$ onto observed sequencing depth of each cell $\ell_1, \ldots, \ell_n$. These models account for the possibility that the sequencing depth detrimentally confounds the differential expression of gene j in varying degrees across all genes. Hence, our modeling of the sequencing depth in (3) is qualitatively similar.

High-level description of the hypothesis test

We provide a high-level description of the hypothesis that eSVD-DE tests, with respect to the hierarchical model prescribed in the aforementioned statistical model. Let $\mathcal{A}$ denote the set of case individuals and $\mathcal{B}$ denote the set of control individuals. Consider a gene $j \in \{1, \ldots, p\}$, assume that all there are parameters $\mu_j^{(\text{case})}$ and $v_j^{(\text{case})}$ that depict the unknown gene expression mean and variance among the case individuals, i.e.,

$$\mu_j^{(s)} \sim \text{Gaussian}\big(\mu_j^{(\text{case})}, v_j^{(\text{case})}\big), \quad \text{for each } s \in \mathcal{A},$$

and likewise for the control individuals for unknown parameters $\mu^{(\text{control})}$ and $\sigma^{(\text{control})}$,

$$\mu_j^{(s)} \sim \text{Gaussian}\big(\mu_j^{(\text{control})}, v_j^{(\text{control})}\big), \quad \text{for each } s \in \mathcal{B}.$$

Each individual then contributes multiple cells. Let $\mathcal{I}(s) \subset \{1, \ldots, n\}$ denote the set of cells that an individual contributes. We model cell i's expression (from individual s, where s refers to "subject") for gene j as

$$\lambda_{ji} \sim F_j^{(s)}, \quad \text{for } i \in \mathcal{I}(s), \tag{5}$$

where $F_j^{(s)}$ is a distribution with mean $\mu_j^{(s)}$. From here, the observed count in cell i for gene j, defined as A_{ji} is related to λ_{ji} through (1).

With this alternative perspective of our statistical model, eSVD-DE is testing between the null and alternative hypotheses,

$$H_{0,j} : \mu_j^{(\text{case})} = \mu_j^{(\text{control})}, \quad \text{and} \quad H_{1,j} : \mu_j^{(\text{case})} \neq \mu_j^{(\text{control})}, \tag{6}$$

which are comparing the pouplation case-individual mean expression to the pouplation control-individual mean expression of gene j.

A few notes are in order:

- **Rationale of against pseudobulk approaches**: The hypothesis test in (6) might suggest a pseudobulk approach (such as used by DESeq2 in our comparisons). However, as we have mentioned in the main text, such pseudobulk methods do not capture variability within an individual (i.e., the variance of the distributions $F_j^{(s)}$). This can affect validity of such pseudobulk strategies, as demonstrated in our null simulations shown later.
- **Rationale of our hypothesis test**: We note that our hypothesis test in (6) relates the population case individual's against the population control individual's mean expression. This is different from the following null hypotheses that are not as apt for testing for differential mean expression for cohort-wide scRNA-seq data:

$$H'_{0,j} : \text{Mean}_{i\in\mathcal{I}(s),s\in\mathcal{A}}\big(\{\lambda_{ij}\}\big) = \text{Mean}_{i\in\mathcal{I}(s),s\in\mathcal{B}}\big(\{\lambda_{ij}\}\big), \tag{7}$$

or

$$H''_{0,j} : \text{Mean}_{i\in\mathcal{I}(s),s\in\mathcal{A}}\big(\{\mu_{ij}\}\big) = \text{Mean}_{i\in\mathcal{I}(s),s\in\mathcal{B}}\big(\{\mu_{ij}\}\big). \tag{8}$$

  We cannot test the null hypothesis in (7) directly since we observe only one observation A_{ji} for each λ_{ji}. By pooling all the information across cells and genes, we can instead estimate μ_{ji}, which is the mean of λ_{ji} based on a low-rank matrix factorization. In contrast, (8) ignores which cells originated from which individual, which makes it undesirable for differential testing for cohort-wide scRNA-seq data. For example, if all the cells from a few case individuals have vastly higher gene expression than all other case individuals' cells, the null hypothesis in (8) might be rejected, but eSVD-DE's null hypothesis in (6) would not be rejected. In such a scenario, we would not want to reject the null hypothesis since the behavior of a small number of case individuals would not be representative of the human population of case individuals.

  In our comparisons, methods like SCTransform test the null hypothesis in (8). Methods like SCTransform face a different computational obstacle. Since these methods regress out the individuals' substantial covariate effects one gene at a time, this regression might be inaccurately estimated due to the sparsity of scRNA-seq data.

- **Distinction between cell-mean and individual-mean**: Observe that even though our hypothesis testing framework in (5) treats all the λ_{ji}'s for the same individual (i.e., $i \in \mathcal{I}(s)$) as i.i.d., eSVD-DE nonetheless models each cell's mean as μ_{ji} in (2). This is to facilitate the matrix factorization framework in order to pool information across cells and genes. As we will discuss later in the model, we estimate $\mu_j^{(s)}$ by averaging the estimated μ_{ji}'s (after taking its posterior distribution to account for model misspecification, see (14)).

We believe it is important to rely on the posterior distribution especially for cohort-wide differential expression testing since single-cell sequencing data is quite sparse, so it's important to have a data-driven procedure to adjust our estimated values of μ_{ji} that balances how sparsely sequenced (or said differently, how much "information") gene j is and how large the covariate-adjusted sequencing depth of cell i is.

- **Covariate effects, sparsity, and overdispersion**: Within this perspective of the model, all the confounding covariate effects is modeled through the covariate-adjusted sequencing depth ℓ_{ji} in (3), which eSVD-DE tries to model appropriately by pooling information across cells and genes. Additionally, we lose power to test for (6) as sparsity increases, which is parameterized by the overdispersion parameter γ_j.

Illustration of how dimension reduction introduces downstream Type-1 inflation in differential expression testing

Here, we provide a simplistic example of how dimension reduction complicates the validity of differential expression testing downstream. Our example relies solely on two genes across six individuals (three cases, three controls), where each gene's expression is Gaussian distributed. To focus on how dimension reduction interacts with downstream testing, we do not involve the count-nature of scRNA-seq data, sparsity of gene expression, or confounding covariate effects in this example.

The simple simulated data is shown in Figure S1A, where Gene 1 is differentially expressed whereas Gene 2 is not (two-sample test p-value of 0.0019 and 0.4723 respectively). Coincidentally, the direction of most variance is also strongly aligned with the axis of Gene 1. Hence, if we project the data onto the leading principal component, our resulting "denoised" dataset is shown in Figure S1B. When the same two-sample test is applied to each gene individually, both genes are differentially expressed (p-value of 0.0019 for both genes). Both genes have the same p-value in this example after projecting onto the first principal component since the "information" in both genes is identical (i.e., perfectly collinear).

This example illustrates the following phenomenon. A dimension reduction (such as projecting data onto its leading principal components) will necessarily introduce correlation among all genes. When true DE genes primarily drive the matrix factorization, many true null genes will appear as statistically significant DE genes downstream after the dimension reduction. In many biological datasets, this is likely the scenario since a few DE genes often have a high signal-to-noise ratio, and the practitioner is interested in finding other DE genes with a weaker signal-to-noise ratio. However, after applying a dimension reduction, most genes will appear to have a DE gene with a weak signal-to-noise ratio.

In our particular example, it was clear that such a dimension reduction was not necessary for our goal of testing for DE genes. However, when there are thousands of genes (instead of two) that are sparsely observed count data (instead of densely observed Gaussian data) that are confounded by various covariates (instead of not having any covariate effects), dimension reduction is likely necessary to pool information across cells and genes. In that setting, the phenomenon we illustrate in

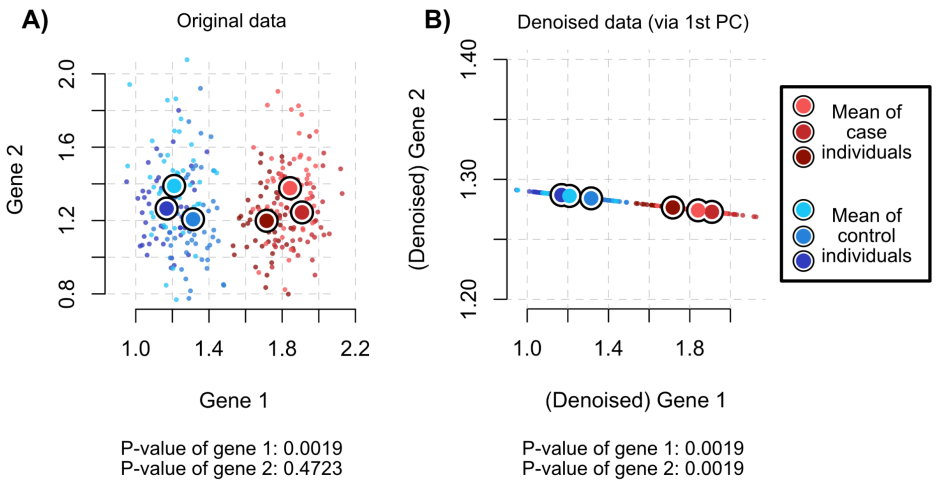

Figure S1 **A** Scatter plot of the simulated dataset with multiple cells originating from six individuals (three cases and three controls) for two genes, where Gene 1 is designed to be a true DE gene (two-sample test p-value of 0.0019) and Gene 2 is not (p-value of 0.4723). **B** Scatter plot of the same dataset after all the cells are projected onto the leading principal component. Now, both genes appear to be statistically significant DE genes (p-values of 0.0019), illustrating how dimension reduction can inflate the downstream Type-1 error.

Figure S1 is likely to happen still but would not be as easily detectable. This illustration highlights the necessity to consider the posterior distribution, which we describe later in (14) as a way to account for model misspecification by balancing how sparsely sequenced gene j is with how large the covariate-adjusted sequencing depth of cell i is.

We note that work using deep variational autoencoders like [16, 17] have developed another way to circumvent this problem. First, the nature of the variational autoencoder relies on simulating Gaussian data after encoding the single-cell gene expression profile (hence, "variational"). This necessarily weakens the correlation among genes despite reducing the dimensionality, in contrast to a matrix-factorization approach like principal component analysis. Second, they advocate adding a "pseudcount" to all the gene expressions, which filters out genes that are deemed as significantly differentially expressed but have a near-zero log-fold change. While both qualities provide a sound solution to perform DE testing after a dimension reduction, this strategy relies on a "black-box" method with numerous tuning parameters to train the deep neural network and to perform the DE test. We instead choose to develop a strategy reliant on the posterior distribution in eSVD-DE as a more transparent and tuning-free strategy.

Implementation and details of eSVD-DE

Initialization

The initialization broadly falls into three steps: 1) initializing the estimate of the coefficients $\widehat{Z} \in \mathbb{R}^{p \times r}$, 2) initializing the estimate of the embeddings for the cells $\widehat{X} \in \mathbb{R}^{n \times k}$ and for the genes $\widehat{Y} \in \mathbb{R}^{p \times k}$, and 3) reparameterizing our estimates $\widehat{X}$, $\widehat{Y}$ and $\widehat{Z}$. We describe each step below.

First, to initialize our estimate $\widehat{Z}$, we fit a Poisson regression with a ridge penalty separately for each cell i. This regresses the covariates C onto $A_{\cdot,i}$, the observed

gene counts for cell i. Specifically, for a small penalty $\tau > 0$, we initialize the jth row of $\widehat{Z}$ as

$$\widehat{Z}_{j,\cdot} = \underset{z \in \mathbb{R}^r}{\operatorname{argmin}} - \Big[\sum_{i=1}^{n} A_{ji} \cdot \big(z^\top C_{i,\cdot}\big) - \exp\big(z^\top C_{i,\cdot}\big) \Big] + \tau \cdot \|z_{-1}\|_2^2, \quad \text{for } j \in \{1, \ldots, p\},$$

where $z_{-1} = (z_2, \ldots, z_r) \in \mathbb{R}^{r-1}$. The ridge penalty omits z_1 because, by definition, $C_{\cdot,1}$ is the all-one vector that represents the intercept. Typically, we set $\tau = 0.01$, a small non-negative value to mitigate multicollinearity issues, similar to other works such as [12].

Second, we initialize our estimates $\widehat{X}$ and $\widehat{Y}$. Consider the matrix,

$$R = \log(A + 1) - \widehat{Z}^\top C \in \mathbb{R}^{p \times n},$$

where $\log(\cdot)$ of a matrix denotes taking the natural logarithm of each matrix entry. Consider the SVD of R,

$$R = UDV^\top,$$

where $U \in \mathbb{R}^{p \times k}$ and $V \in \mathbb{R}^{n \times k}$ are column-wise orthonormal matrices, and D is a diagonal non-negative matrix with decreasing values along the diagonal. We then initialize our estimates of $\widehat{X}$ and $\widehat{Y}$ as

$$\widehat{X} = V\sqrt{D} \in \mathbb{R}^{n \times k}, \quad \text{and} \quad \widehat{Y} = U\sqrt{D} \in \mathbb{R}^{p \times k},$$

where $\sqrt{\cdot}$ of a diagonal non-negative matrix denotes taking the square root of all its diagonal entries. This initialization of $\widehat{X}$ and $\widehat{Y}$ is motivated by the observation that if we model A_{ji} as a Poisson random variable, then $\log(A_{ji} + 1)$ is a crude approximation of the natural parameter μ_{ji}, and $(\widehat{Z}_{j,\cdot})^\top C_{i,\cdot}$ would be an initial estimate of the covariate effect on μ_{ji}.

This step is collectively handled by `eSVD2::initialize_esvd` in our codebase. It relies on the `glmnet::glmnet` function to fit the Poisson ridge regression.

Optimization of the embeddings

The iterative optimization of the embedding broadly falls into two steps to estimate the latent factors of the mean matrix μ defined in (2): 1) updating the estimates of $\widehat{X}$, $\widehat{Y}$, and $\widehat{Z}$, holding the initialized values of $\widehat{Z}_{\cdot,(cc)}$ fixed, and then 2) updating the estimates of $\widehat{X}$, $\widehat{Y}$, and $\widehat{Z}$, allowing the values in $\widehat{Z}_{\cdot,(cc)}$ to be updated. This step takes inspiration from methods such as GLM-PCA [12], ZINB-WaVE [13], and our previous work [18]. We first describe the details of the procedure, followed by its justification.

For a tuning parameter $\tau > 0$, we seek to optimize the following objective function

$$\big\{\widehat{X}, \widehat{Y}, \widehat{Z}\big\} = \underset{X,Y,Z}{\operatorname{argmin}} \Big[-\frac{1}{np} \Big[\sum_{ij} \log P\big(A_{ji} \mid X_{i,\cdot}, Y_{j,\cdot}, C_{i,\cdot}, Z_{j,\cdot}\big) \Big] + \tau \cdot \Big[\|X\|_F^2 + \|Y\|_F^2 + \|Z\|_F^2 \Big]. \tag{9}$$

Let us define the natural parameter,

$$\theta_{ji} = (Y_{j,\cdot})^\top X_{i,\cdot} + (Z_{j,\cdot})^\top C_{i,\cdot}, \quad \text{for all } i \in \{1, \ldots, n\}, j \in \{1, \ldots, p\}.$$

Then, here, $\log P(A_{ji} \mid X_{i,\cdot}, Y_{j,\cdot}, C_{i,\cdot}, Z_{j,\cdot})$ denotes the log-likelihood for a Poisson random variable,

$$\log P\big(A_{ji} \mid X_{i,\cdot}, Y_{j,\cdot}, C_{i,\cdot}, Z_{j,\cdot}\big) = A_{ji} \cdot \theta_{ji} - \exp\big(\theta_{ji}\big).$$

The objective function in (9) is a non-convex, as documented in [18]. This poses challenging optimization considerations. Hence, we first describe an alternating minimization strategy that improves upon previous work computationally. Then, we describe the aforementioned two-step approach.

Our alternating minimization consists of the following steps, starting on an initial estimation of $\widehat{X}^{(0)} = \widehat{X}$, $\widehat{Y}^{(0)} = \widehat{Y}$, and $\widehat{Z}^{(0)} = \widehat{Z}$. This strategy is motivated by the observation that holding Y and Z fixed, the optimization in (9) over X is convex, and vice-versa. Let t denote the iteration counter.

- Optimize the cell embedding,

$$\widehat{X}^{(t+1)} = \operatorname*{argmin}_{X} \Big[-\frac{1}{np} \Big[\sum_{ij} \log P\big(A_{ji} \mid X_{i,\cdot}, \widehat{Y}^{(t)}_{j,\cdot}, C_{i,\cdot}, \widehat{Z}^{(t)}_{j,\cdot}\big) \Big] + \tau \cdot \Big[\|X\|_F^2 \Big]. \tag{10}$$

- Optimize the gene embedding and coefficients

$$\big\{\widehat{Y}^{(t+1)}, \widehat{Z}^{(t+1)}\big\} = \operatorname*{argmin}_{Y,Z} \Big[-\frac{1}{np} \Big[\sum_{ij} \log P\big(A_{ji} \mid \widehat{X}^{(t+1)}_{i,\cdot}, Y_{j,\cdot}, C_{i,\cdot}, Z_{j,\cdot}\big) \Big] + \tau \cdot \Big[\|Y\|_F^2 + \|Z\|_F^2 \Big]. \tag{11}$$

We repeat until convergence and then perform a reparameterization (described in the next section).

The optimizations (10) and (11) are performed using a Newton optimization, which is a second-order method. This specific optimization framework is ideal for solving (10) and (11) for a few reasons:

1 **Parallelization into many low-dimension optimizations**: Both optimizations (10) and (11) decompose into n and p smaller optimization problems respectively. For example, the ith row of $\widehat{X}^{(t+1)}$ is equivalently solved by the optimization

$$\operatorname*{argmin}_{x \in \mathbb{R}^K} \Big[-\frac{1}{np} \sum_{j=1}^{p} \log P\big(A_{ji} \mid x, \widehat{Y}^{(t)}_{j,\cdot}, C_{i,\cdot}, \widehat{Z}^{(t)}_{j,\cdot}\big) \Big] + \tau \cdot \|x\|_2^2. \tag{12}$$

Hence, solving (10) and (11) amounts to a solving n different K-dimensional or p different $(K + r)$-dimensional optimizations respectively. Since we are in a setting where $\max(K, r) \ll \min(n, p)$, it is beneficial to use a second-order

method since the Hessian information can yield faster convergence rates in terms of the number of iterations needed to solve optimizations like (12), and there is not a large computational overhead to compute the Hessians to solve (12) (especially when P is the Poisson distribution, where the Hessians are straight-forward to derive and compute).

2 **No randomness in the optimization**: Since (10) and (11) decompose into n and p smaller optimization problems respectively, these optimizations can be embarrassingly parallelized. Hence, there is no need to consider stochastic optimization schemes. This is appealing as this means different practitioners using our method would necessarily obtain the same resulting fit.

3 **Generalization to other exponential families**: While we focus on specifically solving (10) and (11) for the Poisson distribution, using a Newton optimization is ideal for other exponential-family distributions as well. This is because while other exponential-family distributions like the exponential and Negative Binomial have constraints on the natural parameters θ_{ij}, the gradient and Hessians of these log-likelihoods naturally prevent the optimization iterates from violating these constraints. Hence, our codebase can handle modeling situations beyond this paper's scope.

Our optimization procedure is then the following. First, starting from the current estimates of $\widehat{X}$, $\widehat{Y}$, and $\widehat{Z}$ in the previous step ("Initialization"). we updating the estimates of $\widehat{X}$, $\widehat{Y}$, and $\widehat{Z}$ via alternating minimization holding the initialized values of $\widehat{Z}_{\cdot,(\mathrm{cc})}$ fixed (what we'll call the "Phase one optimization"). After convergence, we then perform a second round of alternating minimization to update the estimates of $\widehat{X}$, $\widehat{Y}$, and $\widehat{Z}$, allowing the values in $\widehat{Z}_{\cdot,(\mathrm{cc})}$ to be updated (what we'll call the "Phase two optimization"). (Here, we omit the superscript "(t)" for notational simplicity.) We do these two phases of optimization since empirically, we have observed more stable behavior and a better final objective value doing two rounds of optimization (compared to optimizing all of X, Y, and Z, including $Z_{\cdot,(\mathrm{cc})}$ from the start). This becomes imperative since $Z_{\cdot,(\mathrm{cc})}$ will play a more pronounced role in the remainder of eSVD-DE compared to all the other covariates in Z, because is it part of the "signal" that we wish to estimate and is not a "confounder." This is inspired by theoretical results regarding warm-starting the non-convex optimization [53, 54]. This step is collectively handled by `eSVD2::opt_esvd` in our codebase.

Reparameterizing the matrix factorization

After completing either of the two phases mentioned above, we apply the following reparameterization procedure. The goal of this reparameterization is to ensure identifiability since, a priori, there could be multiple estimates $\{\widehat{X}, \widehat{Y}, \widehat{Z}\}$ that have the same predictive power but offer different interpretations of the data. For instance, without such a reparameterization, there could be many columns in $\widehat{X}$ that are correlated with other columns in $\widehat{X}$ or columns in the covariate matrix C, which would obfuscate interpreting different axes of variation. Hence, our reparameterization procedure ensures orthogonality among our estimated matrix factorization to resolve this identifiability concern.

When describing our reparameterization procedure, for notational simplicity, we let $\{\widehat{X}, \widehat{Y}, \widehat{Z}\}$ denote $\{\widehat{X}^{(T)}, \widehat{Y}^{(T)}, \widehat{Z}^{(T)}\}$, which is the final estimate after T itera-

tions for either the Phase one or two optimizations. We also let the "$a \leftarrow b$" notation denote setting the variable a to be the value in variable b.

The reparameterization procedure operates in two steps. The first step ensures that $\widehat{X}$ is orthogonal to C (which would result in adjusting $\widehat{X}$ and $\widehat{Z}$). This ensures that $\widehat{X}$ captures variability that is not explained by C. The second step ensures that both $\widehat{X}$ and $\widehat{Y}$ have orthogonal columns (which would result in adjusting $\widehat{X}$ and $\widehat{Y}$). This ensures that each respective latent dimension in both $\widehat{X}$ and $\widehat{Y}$ are identifiable.

- **Step 1**: We first perform a regression of each latent dimension of $\widehat{X}$ onto C. That is, for each latent dimension $d \in \{1, \ldots, k\}$,

  $$\widehat{X}_{\cdot,d} = C^\top \beta + \epsilon,$$

  where $\beta \in \mathbb{R}^r$ and $\epsilon \in \mathbb{R}^n$ are the temporary variables to denote the coefficients for the covariates and residuals respectively. For this latent dimension d, we then perform the following update for each gene $j \in \{1, \ldots, p\}$,

  $$\widehat{Z}_{j,\cdot} \leftarrow \widehat{Z}_{j,\cdot} + \beta \cdot \widehat{Y}_{j,d}, \quad \text{and} \quad \widehat{X}_{\cdot,d} \leftarrow \epsilon.$$

  It can be seen that after performing this update for every latent dimension $d \in \{1, \ldots, k\}$, $\widehat{X}$ is orthogonal to C (i.e., $\widehat{X}^\top C = 0$) even though the predictive power of our factorization (i.e., $\widehat{Y}^\top \widehat{X} + \widehat{Z}^\top C$) did not change.
- **Step 2**: We next perform a linear transformation on $\widehat{X}$ and $\widehat{Y}$. Specifically, using the details in [18], let $R = \widehat{Y}^\top \widehat{X} \in \mathbb{R}^{p \times n}$, with an SVD of $R = UDV^\top$. Then,

  $$\widehat{X} \leftarrow \left(\frac{n}{p}\right)^{1/4} V\sqrt{D}, \quad \text{and} \quad \widehat{Y} \leftarrow \left(\frac{p}{n}\right)^{1/4} U\sqrt{D}.$$

  It can be seen that after performing this update that both $\widehat{X}^\top \widehat{X}/n$ and $\widehat{Y}^\top \widehat{Y}/p$ are diagonal matrices and are equal even though the predictive power of our factorization (i.e., $\widehat{Y}^\top \widehat{X} + \widehat{Z}^\top C$) did not change. This ensures identifability of $\widehat{X}$ and $\widehat{Y}$.

Estimating overdispersion parameter

We estimate the overdispersion parameter $\gamma_1, \ldots, \gamma_p > 0$, one for each gene. Following our model of the covariate-adjusted sequencing depth (3), we estimate the covariate-adjusted sequencing depth of each cell as

$$\widehat{\ell}_{ji} = \exp\left((\widehat{Z}_{j,\text{-(cc)}})^\top C_{i,\text{-(cc)}}\right), \quad \text{for all } i \in \{1, \ldots, n\}, j \in \{1, \ldots, p\}. \tag{13}$$

Likewise, the estimate of the mean parameter is

$$\widehat{\mu}_{ji} = \exp\left((\widehat{Y}_{j,\cdot})^\top \widehat{X}_{i,\cdot} + \widehat{Z}_{j,(\text{cc})} \cdot C_{i,(\text{cc})}\right).$$

Then, using a plug-in estimate of the maximum likelihood of the model in (1), we can derive that the estimate of γ_j is

$$\widehat{\gamma}_j = \max_{\gamma} \sum_{i=1}^{n} A_{ji} \log \widehat{\ell}_i + \widehat{\mu}_{ji}\gamma \log \gamma - \log \Gamma(\gamma \widehat{\mu}_{ji}) + \log \Gamma(A_{ji} + \gamma \widehat{\mu}_{ji}) \\ - (A_{ji} + \gamma \widehat{\mu}_{ji}) \log(\widehat{\ell}_i + \gamma),$$

where $\Gamma(\cdot)$ is the Gamma function. We estimate this via Newton's method, and this step is primarily handled by `eSVD2::estimate_nuisance` in our codebase.

Computing the posterior distribution

To account for possible model misspecification, we estimate the posterior distribution of the mean and variance of each gene j's expression in cell i. This is based on the model in (1), where we can leverage the fact that λ_{ji} conditioned on A_{ji} follows a Negative Binomial distribution. Specifically, based on the posterior distribution derived in (4), we compute

$$\widehat{\mu}_{ji}^{(\text{post})} = \frac{\widehat{\mu}_{ji}/\widehat{\gamma}_j + A_{ji}}{1/\widehat{\gamma}_j + \widehat{\ell}_{ji}}, \quad \text{and} \quad \widehat{v}_{ji}^{(\text{post})} = \frac{\widehat{\mu}_{ji}/\widehat{\gamma}_j + A_{ji}}{(1/\widehat{\gamma}_j + \widehat{\ell}_{ji})^2}, \tag{14}$$

where $\widehat{\mu}_{ji}^{(\text{post})}$ and $\widehat{v}_{ji}^{(\text{post})}$ is the posterior mean and variance of $\lambda_{ji}|A_{ji}$ for gene j in cell i respectively. Other work have used similar modeling based on the posterior distribution for single-cell RNA-seq data [22], but not for the purposes of DE testing. This step is primarily handled by `eSVD2::compute_posterior` in our codebase.

Computing the test statistic

In this step, we compute the test statistic, accounting that multiple cells originate from a particular individual. This broadly falls into two steps: 1) computing the mean and variance for a gene among an individual, and 2) computing the test statistic among all the case and control individuals.

With the posterior distribution in (14), we can compute the expected posterior mean and variance for a particular gene j in individual s. First, we aggregate among the cells for each individual. We assume that after averaging among all the cells from an individual, the resulting distribution can be reasonably approximated by a Gaussian with mean $\widehat{\mu}_j^{(s)}$ and variance $\widehat{v}_j^{(s)}$. Hence, let $\mathcal{I}(s) \subset \{1, \ldots, n\}$ denote the set of cells originating from individual s. Then, via large-sample average,

$$\widehat{\mu}_j^{(s)} = \frac{\sum_{i \in \mathcal{I}(s)} \widehat{\mu}_{ji}^{(\text{post})}}{|\mathcal{I}(s)|}, \quad \text{and} \quad \widehat{v}_j^{(s)} = \frac{\sum_{i \in \mathcal{I}(s)} \widehat{v}_{ji}^{(\text{post})}}{|\mathcal{I}(s)|},$$

where $\widehat{\mu}_{js}$ and $\widehat{v}_{js}$ is the posterior mean and variance for gene j among all the cells in individual s respectively.

Next, we aggregate among individuals. Specifically, among all the case individuals, we can think of the expression of gene j as a Gaussian mixture among the individuals. Let $\mathcal{A}$ denote the set of case individuals. We then summarize the Gaussian

mixture with one Gaussian,

$$\widehat{\mu}_j^{(\text{case})} = \frac{\sum_{s\in\mathcal{A}} \widehat{\mu}_j^{(s)}}{|\mathcal{A}|}, \quad \text{and} \quad \widehat{v}_j^{(\text{case})} = \frac{\sum_{s\in\mathcal{A}} \widehat{v}_j^{(s)}}{|\mathcal{A}|} + \frac{\sum_{s\in\mathcal{A}} (\widehat{\mu}_j^{(s)})^2}{|\mathcal{A}|} - \Big(\frac{\sum_{s\in\mathcal{A}} \widehat{\mu}_j^{(s)}}{|\mathcal{A}|}\Big)^2.$$

We do an analogous calculation among controls, letting $\mathcal{B}$ denote all the control individuals.

Lastly, we do a two-sample T-test with unequal variances, treating the distribution among cases as a Gaussian with mean $\widehat{\mu}_j^{(\text{case})}$ and variance $\widehat{v}_j^{(\text{case})}$ and the distribution among controls as an analogous Gaussian. Our test statistic for gene j is

$$\widehat{T}_j = \frac{\widehat{\mu}_j^{(\text{case})} - \widehat{\mu}_j^{(\text{control})}}{\sqrt{\frac{1}{|\mathcal{A}|} \cdot \widehat{v}_j^{(\text{case})} + \frac{1}{|\mathcal{B}|} \cdot \widehat{v}_j^{(\text{control})}}}. \tag{15}$$

In our codebase, this step is primarily handled by `eSVD2::compute_test_statistic`.

Performing multiple testing correction

We use the multiple testing procedure based on the empirical null developed in [55] to handle the multiple testing corrections. This framework is also used in other methods designed to test for differential expression from scRNA-seq data (albeit not from cohort data), such as iDEA [56] and SifiNet [57]. We know that biologically, most genes are not directly related to the disease or disorder at the human population level; hence, most genes should be deemed insignificant. The empirical null, where the appropriate null distribution is learned from the data, is a suitable framework to ensure this behavior.

Briefly, we first estimate the degree-of-freedom of each gene $j \in \{1, \ldots, p\}$ (used for T-tests for unequal variances) by

$$\widehat{\text{df}}_j = \frac{\Big(\frac{1}{|\mathcal{A}|} \cdot \widehat{v}_j^{(\text{case})} + \frac{1}{|\mathcal{B}|} \cdot \widehat{v}_j^{(\text{control})}\Big)^2}{\Big(\frac{1}{|\mathcal{A}|} \cdot \widehat{v}_j^{(\text{case})}\Big)^2 / (|\mathcal{A}| - 1) + \Big(\frac{1}{|\mathcal{B}|} \cdot \widehat{v}_j^{(\text{control})}\Big)^2 / (|\mathcal{B}| - 1)},$$

and convert all the test statistics $\widehat{T}_j$'s into a Z-score via,

$$\widehat{Z}_j = \Phi^{-1}\Big(F_{\widehat{\text{df}}_j}(\widehat{T}_j)\Big),$$

where $F_b(a)$ is the CDF of the a t-distribution with degree freedom $b > 0$ evaluated at a, and $\Phi^{-1}(a)$ is the quantile function of a standard Gaussian evaluated at $a \in (0, 1)$.

To identify the DE genes, we use the `locfdr::locfdr` function to estimate the empirical null distribution's mean and standard deviation. We then compute the p-value of each gene based on the (two-sided) tail area of a Gaussian distribution with that empirical null's mean and standard deviation. We can compute the negative $\log_{10}$ p-value for visualization purposes in a volcano plot (as in Figures 3 and 4). To perform multiple testing corrections, we apply `stats::p.adjust` on these p-values

via Benjamini-Hochberg. All the genes with an empirical FDR of less than 0.05 are then deemed to be DE genes.

This step is collectively handled by `eSVD2::compute_df`, `eSVD2::compute_test_statistic`, and `eSVD2::compute_pvalue` in our codebase.

Null simulation

We perform the following simulation to demonstrate that eSVD-DE properly controls the Type-1 error. To effectively illustrate this, we construct a simulation where out of the $p = 1000$ simulated genes, 10 of which are truly differentially expressed. The remaining 990 genes are null, and we are interested if the p-values for these 990 genes are uniformly distributed in 0 and 1. We then contrast our method's results with other methods commonly used to analyze cohort-wide single-cell data.

Simulation setup. We simulate data among $p = 1000$ genes across 10 case individuals and 10 control individuals where each individual contributes 100 cells, forming a combined $n = 2000$ cells. The individual covariates we consider are sex (categorical, 5 males and 5 females in each case and control) and age (numerical, drawn from a Gaussian distribution with a mean of 30 and standard deviation of 5). These are encoded to the covariate matrix $C \in \mathbb{R}^{n \times 5}$, where the 5 covariates are the intercept (i.e., a column of all 1's), the log sequencing depth (where it temporarily set to all 0's), and the case-control status, the sex, and age of the cell's individual.

We first describe the generation of the true natural parameter matrix. For the 1000 genes, as mentioned before, the first 10 genes are truly differentially expressed. We simulate the natural parameter matrix $\Theta^* \in \mathbb{R}^{p \times n}$ gene-by-gene, where each gene falls in one of three categories:

1. **True DE gene**: For genes $j \in \{1, 2, \ldots, 10\}$, the mean value among all the case individuals is uniformly chosen between 0.5 and 3. The mean value among all the control individuals is set to be exactly 1 larger or smaller (randomly chosen) than the mean value among all the case individuals. For a group (i.e., either the cases or controls), the specific individual's mean value is sampled to be normally distributed around their respective group's mean with a standard deviation of 0.1. Then, for an individual, all cells originating within an individual are sampled to be normally distributed around the individual's mean with a standard deviation of 0.1.

   The result is a substantial difference between all the case and control individuals' values for this gene j. Figure S2(A,D) shows an example of such a gene.
2. **Null gene with imbalanced variance**: For genes $j \in \{11, 13, \ldots, 999\}$, the mean value among all individuals is uniformly chosen between -0.5 and 1. For both cases and control individuals, the specific individual's mean value is sampled to be normally distributed around their respective group's mean with a standard deviation of 0.1. Then, one group (i.e., the case or control, randomly chosen) is selected so all the individuals in that group have a within-individual standard deviation set to be 0.75 (i.e., high within-individual variation). In contrast, the other group has all the individuals have a within-individual standard deviation set to be 0.1 (i.e., low within-individual variation). Then, for

an individual, all cells originating within an individual are sampled to be normally distributed around the individual's mean with their prescribed standard deviation.

The result is that even though there is no difference in mean between all the cases from all the controls in aggregate for this gene j, the within-individual variation in one group is substantially higher than the other group. Figure S2(B,E) shows an example of such a gene.

3 **Null gene interleaved between case and control**: For genes $j \in \{12, 14, \ldots, 1000\}$, the mean value among all individuals is uniformly chosen between 0.5 and 2. For all individuals, the specific individual's mean value is sampled to be normally distributed around their respective group's mean with a standard deviation randomly chosen between 0.1 and 0.2. Then, for an individual, all cells originating within an individual are sampled to be normally distributed around the individual's mean with a standard deviation randomly chosen between 0.1 and 0.2.

The result is that there is no difference between case and control individuals for this gene j. Figure S2(C,F) shows an example of such a gene.

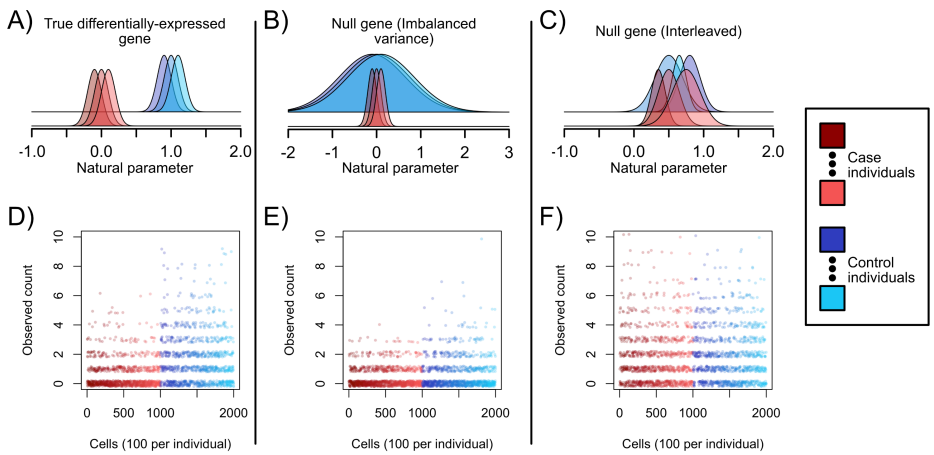

Figure S2 **A** Density of values for a true DE gene, where there is a significant difference in means between the population case individuals and population control individuals, even when accounting for the within-individual variance. **B** Density of values for a null gene where there is a much higher within-individual variance among the control individuals than the control individuals, but the population mean among the case individuals' means is equal to the population mean among the case individuals' means (i.e., null gene with imbalanced variance). **C** Density of values for a null gene where the distribution of any individual's cells' expression is drawn from the same distribution (i.e., interleaved null gene). **D** Observed counts for a true DE gene across the six individuals when there are no confounding covariate effects, displaying a clear higher expression among the control individuals than the case individuals. **E** Observed counts for a null gene with imbalanced variance across the six individuals when there are no confounding covariate effects, displaying a larger variance among the control individuals than the case individuals. **F** Observed counts for an interleaved null gene, where both case and control genes have an unobservable difference in mean or variance. In all six figures, the simulated data displays the gene expression from three case individuals (in different shades of red) and three control individuals (in different shades of blue).

Then, we threshold the entries in Θ^* to not be larger than $\log(100)$, then uniformly shift all the entries down to encourage more sparsity in the generated observed matrix downstream.

We then describe how to introduce the sequencing-depth effects based on the covariates. Due to how the covariate matrix C is constructed, the only relevant columns of the coefficient matrix $Z^* \in \mathbb{R}^{p\times 5}$ that need to be discussed are the coefficients for the intercept (which are set to all 0), the sex (which are sampled to be normally distributed with mean 0 and standard deviation 0.2), and the age (which are sampled to be normally distributed with mean 0 and standard deviation 0.5). Then, we construct

$$\ell^*_{ji} = \exp\left((Z^*_{j,\cdot})^\top C_{i,\cdot}\right).$$

We then describe the generation of the overdispersion parameters. The overdispersion parameter γ^*_j is set to be 1 for all truly DE genes (i.e., $j \in \{1, \ldots, 10\}$), and then uniformly randomly sampled among 0.1, 1, and 10 for all remaining 990 null genes.

Lastly, we describe the generation of the observed simulated data. Equipped with all the above ingredients, we simulated data according to the hierarchical model in (1). Specifically, letting the mean value of gene j in cell i be

$$\mu^*_{ji} = \exp\left(\Theta^*_{ji}\right),$$

we sample the observed data matrix $A \in \mathbb{Z}_+^{p\times n}$ to be,

$$\left(A_{ji}|\lambda^*_{ji}\right) \sim \text{Poisson}(\ell^*_{ji}\cdot\lambda^*_{ji}), \quad \text{and} \quad \lambda^*_{ji} \sim \text{Gamma}(\alpha = \mu^*_{ji}/\gamma^*_j\ ;\ \beta = 1/\gamma^*_j),$$

The resulting dataset is roughly 60% sparse (i.e., only 40% of the entries are non-zero). We then appropriately set the values in $C_{\cdot,(\text{lib})}$ to be the log sequencing depth of the observed data.

We rationalize our design of the simulated data for the following reasons:

- **Clear definition of DE gene**: By the way we generated the values in Θ^*, it is clear that in population, only genes $j \in \{1, \ldots, 10\}$ are differentially expressed because all the other genes have the same group-wise mean between cases and controls.
- **Demonstration of scenario where pseudobulk is not ideal**: Due to the null genes with imbalanced variances (i.e., genes $j \in \{11, 13, \ldots, 999\}$), the higher variation within-individual in one group versus the others means that if only pseudobulk samples were considered, the analysis loses information on how variable the cells are within an individual. Since the observed values in A are only non-negative integers, a higher variance within an individual typically results in a higher observed mean. This means in this analysis, the pseudobulk samples could mistake cells with a higher within-individual variance to have a higher within-individual mean. In contrast, an analysis on the single-cell level like eSVD-DE can retain the information that the variation within the individual is large for downstream test statistic calculations.
- **Demonstration of the necessity of pooling information in the presence of sparsity**: Lastly, the sparsity of each gene means that regressing out the covariate effect gene-by-gene is likely to result in a suboptimal removal of covariate effects.

Methods used. We compare our eSVD-DE method against four other methods described previously in the main text. These are: 1) DESeq2, as a prototypical method that illustrates analyzing the data via pseudobulk samples, 2) MAST, as a prototypical method that illustrates analyzing the data via mixed-effect models, 3) SCTransform followed by Wilcoxon test, as a prototypical method that illustrates analyzing the data when ignoring which cells originate from which individuals, and 4) GLM-PCA followed by the Wilcoxon rank sum test, which prototypical method that illustrates analyzing the data using existing dimension-reduction pipelines. Notably, we are *not* critiquing any particular method specifically, but instead the overarching strategy of either regression out the covariate effects one gene at a time or using a dimension-reduction method followed by a Wilcoxon rank sum test respectively.

The goal is 1) to ensure that the 10 true DE genes are indeed assessed to be differentially expressed, but more importantly, 2) to verify that the remaining 990 null genes have a uniform distribution of p-values.

Results. When applying all 5 methods to the data, all 5 methods consistently rejected the null hypothesis for the 10 true DE genes (across multiple trials). This ensured that each method had sufficient detection power.

However, when analyzing the p-values for the remaining 990 null genes, each method varied in the amount of Type-1 error. The QQ plots of the null genes' p-values are shown in Figure S3 (i.e., both imbalanced variance and interleaved). We see that eSVD-DE correctly has no Type-1 inflation among the null genes due to its pooling of information across cells and genes in the matrix factorization framework but also the posterior correction to avoid model misspecification. DESeq2 and MAST were the next two methods with the least Type-1 error, which reflects why pseudobulk or mixed-effect models are often deployed to analyze cohort-wide scRNA-seq data currently [6, 4, 21]. DESeq2's Type-1 inflation occurred specifically the genes with imbalanced variance. This is expected, as pseudobulk methods do not capture the within-individual variability. While MAST's Type-1 error stemmed from its inaccuracy to regress out the covariate effect (due to its regression gene-by-gene), it performed better than SCTransform and GLM-PCA due to its mixed-effect nature. Lastly, SCTransform and GLM-PCA inflated Type-1 errors dramatically. Importantly, we are not critiquing these methods specifically but rather the pitfalls of using these overarching strategies when testing for DE in cohort-wide scRNA-seq data. The Type-1 inflation is due to its inaccuracy to regress out the covariate effect and/or testing for DE on the cell-level instead of at the individual-level, as discussed throughout the main text and appendix.

Details of power simulation

We now describe how the simulated data in Figure 2 was generated. The complexity of our data generation scheme stems from the desire to simultaneously estimate a gene's "true" log-fold change between cases and controls (excluding confounding effects) while also having genes of varying log-fold changes. We specifically do not want to simulate genes with explicitly no "true" log-fold change in this simulation. Instead, we wish to see if eSVD-DE's estimated log-fold change of a gene is strongly

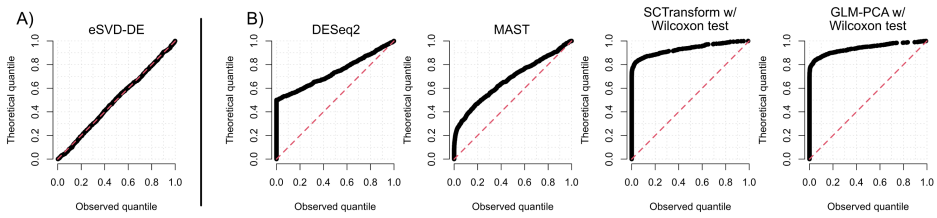

Figure S3 **A** QQ-plot of the quantiles of the p-values for the null genes (theoretical vs. observed) for eSVD-DE on the null simulation in one trial. **B** For the same trial, the QQ-plots of the p-values for DESeq2, MAST, SCTransform (followed by a Wilcoxon rank sum test) and GLM-PCA (followed by a Wilcoxon rank sum test). In every plot, the red dotted line denotes the $x = y$ line. A method with no Type-1 inflation would show that the theoretical equal the observed quantiles among the null genes, which is only achieved by eSVD-DE in our null simulation.

correlated with the "true" log-fold change of a gene. We also wish to make the data generation process complex so that it is misspecified with respect to eSVD-DE's assumed model. This quality helps to ensure that no method, not even eSVD-DE, has an unfair advantage when benchmarking eSVD-DE against other methods.

Since the data generation method is quite involved, please see the codebase for details. Below, we summarize the main aspects of the procedure.

Recall the model in (1). The first phase of data generation is to generate the cell embedding $X \in \mathbb{R}^{n \times k}$, the gene embedding $Y \in \mathbb{R}^{p \times k}$, the cell's covariate matrix $C \in \mathbb{R}^{n \times r}$, the coefficient matrix $Z \in \mathbb{R}^{p \times r}$, and the overdispersion vector $\gamma \in \mathbb{R}^p_+$. To do this, we simulate a population of 20 individuals (10 cases and 10 controls), each contributing 250 cells (meaning $n = 5000$ cells). The cell embedding X is generated from k-dimensional multivariate Gaussians with non-trivial correlation, where $k = 10$. On the other hand, the gene embedding Y by assigning genes to a weighted combination of 4 "topics" or no topics at all. For the genes assigned to the former category, the latent embedding for the gene is drawn from a simplex with 4 points (in a k-dimensional space). The latent embedding is drawn from a multivariate Gaussian centered at 0 for genes in the latter category. In total, there are $p = 700$ genes, where 200 of the genes are not assigned to any topic. For the covariate matrix C, the case-control status is set accordingly so the first 10 individuals are cases, and the last 10 individuals are controls. There is also a column for the intercept, a column for the log sequencing depth (which is set to be 0 for now), as well as the covariates for age (normally distributed), sex (boolean, equally distributed between cases and controls), and tobacco usage (boolean, correlated with the case-control status). Hence, $r = 6$. The coefficient matrix Z is constructed column-by-column, where the coefficient for a gene is set according to which topic the gene was assigned to, and different topics have different distributions of the coefficient value. The coefficient for the log sequencing depth is set to be all 1's. Lastly, the nuisance parameter, taking values from 0.1, 1, and 100, is also set according to the topic the gene was assigned to, where, in general, if a gene has a larger case-control coefficient, we set the gene to be less overdispersed.

Equipped with all the components of the model, we can generate the low-dimensional mean matrix $\mu \in \mathbb{R}^{p \times n}$ according to (2). However, to make the model misspecified with respect to the eSVD-DE, we perform two additional modifications

to the generated low-dimensional mean matrix μ. First, depending on which genes had a large case-control coefficient in magnitude, we slightly shift all the 250 cells for each of the 20 individuals by a constant amount. This way, there is more variation among these genes that is not solely captured by the boolean case-control status vector in C. Second, for many of the genes with a small case-control coefficient, we artificially shrink the expression for that gene to become more correlated with the genes with a large case-control coefficient. This way, there is more correlation among the p genes. We then generate the matrix $\lambda \in \mathbb{R}^{p \times n}$ and the observed count matrix $A \in \{0, 1, \ldots, \}^{p \times n}$ from the modified mean matrix μ and the overdispersion vector γ, following (1).

Once the observed count matrix has been generated, we can compute the observed log sequencing depth of each cell and set the column accordingly to $C_{\cdot,(\text{lib})}$. To assess the "true" gene expression not impacted by the confounding covariates, we regress μ onto all the columns of C aside from the column for the case-control status, $C_{\cdot,(\text{cc})}$, and inspect the residuals. For these residuals, we can compute the "true" test statistic of each gene unaffected by the confounding covariates using calculations akin to those in the "Computing the test statistic" and "Performing multiple testing correction" subsections above. This lets us know the true log-fold change between case and control individuals for each gene. Also, we can compute the "true" FDR of each gene, and we deem all genes with a "true" FDR of less than 0.05 as a "true" DE gene.

Additional details of data preprocessing and analysis

Details of data preprocessing

- **Adams et al., and Habermann et al.**: We select T cells for our analysis (i.e., `Manuscript_Identity=="T"`) and the individuals either in the control set or with IPF (i.e., `Disease_Identity %in% c("Control", "IPF")`), and preprocess the scRNA-seq data using `Seurat::NormalizeData` (with `normalization.method = "LogNormalize"`), `Seurat::FindVariableFeatures` (with `selection.method = "vst"` and `nfeatures = 5000`).

  We select T cells for our analysis (i.e., `celltype=="T Cells"`) and the individuals either in the control set or with IPF (i.e., `Diagnosis %in% c("Control", "IPF")`). To ensure that the analysis between both datasets is comparable, we set the highly variable genes in both datasets to be the 5000 highly variable genes (HVG) in either the Adams and Habermann dataset (after using `Seurat::NormalizeData` with `normalization.method = "LogNormalize"` as well as `Seurat::FindVariableFeatures` with `selection.method = "vst"` and `nfeatures = 5000`), as well as the DE genes found for T cells in either [6] or [19], housekeeping genes and cell-cycling genes (from the `Seurat` package, in `cc.genes$s.genes` and `cc.genes$g2m.genes`). This resulted in the 8909 genes used in our analysis for each dataset. Importantly, the usage of `Seurat::NormalizeData` is soley for defining the highly variable genes, and the eSVD-DE is applied to the original count data.

  For analyzing the Adams dataset with eSVD-DE, we use the covariates `"Disease_Identity"` (which dictate the case-control status), `"Subject_Identity"`

(which dictate which cells originated from which individual), "Gender", "Tobacco", "percent.mt" and "Age". When initializing the eSVD-DE fit, we use $k = 15$ latent dimensional embedding for X and Y.

For analyzing the Adams dataset with eSVD-DE, we use the covariates "Diagnosis" (which dictate the case-control status), "Sample_Name" (which dictate which cells originated from which individual), "Gender", "Tobacco", "percent.mt" and "Age". When initializing the eSVD-DE fit, we use $k = 15$ latent dimensional embedding for X and Y, and

- **Smillie et al.**: We focus on the cell types encoded in Celltype to be "Cycling TA", "Enterocyte Progenitors", "TA 1", or "TA 2". We remove cells that have one of the highest gene expressions in too many genes, as well as genes that are too sparsely observed. To select the genes for our analysis, we use the highly variable genes (after using Seurat::NormalizeData with normalization.method = "LogNormalize" as well as Seurat::FindVariableFeatures (with selection.method = "vst" and nfeatures = 5000), as well as any DE genes found by [20] (when comparing inflamed to healthy, non-inflamed to healthy, or inflamed to non-inflamed), the housekeeping genes, and the cell-cycling genes. Importantly, the usage of Seurat::NormalizeData is soley for defining the highly variable genes, and the eSVD-DE is applied to the original count data. Finally, we keep the cells whose variance across all the selected genes is not too low.

  When analyzing any cell type using eSVD-DE, we first compare the healthy individuals (i.e., Subject_Disease == "HC" and Sample_Health == "Healthy") to case individuals with a tissue section that is inflamed (i.e., Subject_Disease == "Colitis" and Sample_Health == "Inflamed"). We perform a separate analysis that compares the healthy individuals to case individuals with a tissue section that is non-inflamed (i.e., Subject_Disease == "Colitis" and Sample_Health == "Non-inflamed"). To ensure that our analysis of healthy vs. inflamed compared to our analysis of healthy vs. non-inflamed shows biologically meaningful correlation, we split the healthy individuals into two groups. We use only one group of healthy individuals in each analysis.

  For analyzing the any cell type, we use the covariates "Subject_Disease" (which dictate the case-control status), "Sample" (which dictate which cells originated from which individual), "Subject_Gender", "Subject_Smoking", "Subject_Location", and "percent.mt". When initializing the eSVD-DE fit, we use $k = 15$ latent dimensional embedding for X and Y.

- **Velmeshev et al.**: We focus on the cell types encoded in celltype to be "AST-PP", "Endothelial", "IN-SST", "IN-VIP", "L2/3", "L4", "L5/6", "L5/6-CC", "Microglia", "Oligodendrocytes", and "OPC". We remove cells that have one of the highest gene expressions in too many genes, as well as genes that are too sparsely observed. To select the genes for our analysis, we use the highly variable genes (after using Seurat::NormalizeData with normalization.method = "LogNormalize" as well as Seurat::FindVariableFeatures

(with `selection.method = "vst"` and `nfeatures = 5000`), as well as any DE genes found by [21], the SFARI genes, the housekeeping genes, and the cell-cycling genes. Importantly, the usage of `Seurat::NormalizeData` is soley for defining the highly variable genes, and the eSVD-DE is applied to the original count data. Finally, we keep the cells whose variance across all the selected genes is not too low.

For analyzing the any cell type for either of the comparisons, we use the covariates `"diagnosis"` (which dictate the case-control status), `"individual"` (which dictate which cells originated from which individual), `"region"`, `"age"`, `"sex"`, `"Seqbatch"`, and `"Capbatch"`. When initializing the eSVD-DE fit, we use $k = 30$ latent dimensional embedding for X and Y.

In our application of eSVD-DE, we set the regularization parameter $\tau = 0.1$ for the initialization as well as the alternative minimization.

Details of visualizing the embedding

To visualize the embeddings (either of the principal components of the log-normalized scRNA-seq data or of the low-dimensional embedding X estimated by eSVD-DE, throughout Figures 2 through 5), we use the `dimRed::embed` function (with `.method="Isomap"` and `knn=30`).

Details of exploring the relationship between expression and sequencing depth

To explore the relationship between the gene expression and sequencing depth (as demonstrated in Figure 3), we deploy a procedure similar to the one used in [39].

Specifically, for the "Before eSVD-DE" analysis, we first cluster all the genes into 6 clusters based on the logarithm of the total expression across all cells. Then, for each gene separately, we regress the gene's expression (across all the cells) onto the log sequencing depth (across all the cells) via a kernel nonparametric regression using the `npregfast::frfast` function. After performing this regression, we compute the predicted gene expression for different log sequencing depths. Finally, for all the genes within one of the 6 clusters, we compute the median as well as the upper and lower quantiles for the predicted gene expression for different log sequencing depths. This is visualized in Figure 3C.

For the "After eSVD-DE" analysis, we apply the same procedure, except after fitting the eSVD-DE, we regress the denoised gene's expression, not including the covariate-adjusted sequencing depth (across all n cells), i.e., for gene j,

$$\left\{\widehat{\mu}_{ji}\right\}_{i=1}^{n} = \left\{ \exp\left((\widehat{Y}_{j,\cdot})^\top (\widehat{X}_{i,\cdot}) + \widehat{Z}_{j,(\mathrm{cc})} \cdot C_{i,(\mathrm{cc})} \right) \right\}_{i=1}^{n},$$

onto the log sequencing depth (across all the cells) via a nonparametric regression, and then proceed with the remainder of the analysis similar to the "Before eSVD-DE" analysis. This is visualized in Figure 3D.

Details of DESeq2

For all analyses using DESeq2 [8], we first construct pseudobulk samples for each individual, whereby for each gene, we sum the counts across all the cells from that

individual. If there is no variation among a covariate among all the cells for this individual, we set the covariate for the pseudobulk to be equal to the unique covariate from the individual. We use the mean value for covariates such as `percent.mt` that vary across the cells. Then, we use `DESeq2::DESeqDataSetFromMatrix` and `DESeq2::DESeq` to perform the DE test, whereby we use a Benjamini-Hochberg multiple-testing correction afterward.

Details of MAST

We first log-normalize the RNA counts for all analyses using MAST [10]. Then, we use the `MAST::FromMatrix` function (setting `cData` accordingly based on the cells' covariates) and the `MAST::zlm` function to fit the mixed-effects model, using a random effect for the individual (with `method = "glmer"` and `ebayes = FALSE`), followed by the `MAST::lrTest` function to perform the DE test, whereby we use a Benjamini-Hochberg multiple-testing correction afterward.

Details of SCTransform

For all analyses using SCTransform [39], we use the `Seurat::SCTransform` (with `method = "glmGamPoi"`) and set the `vars.to.regress` argument to be all the relevant covariates (notably, not including any one-hot encoding vectors that depict which cell originated from which individual). Then, we use `Seurat::FindMarkers` function (with `test.use = "wilcox"`, `logfc.threshold = 0`, and `min.pct = 0`) to perform the DE test, whereby we use a Benjamini-Hochberg multiple-testing correction afterward.

Details of GLM-PCA

For all analyses using GLM-PCA [12], we use the `glmpca::glmpca` (with `fam = "poi"`, `minibatch = "stochastic"`) with an equal number of latent dimensions as the corresponding eSVD-DE analysis, and setting `X` to be the relevant covariates (excluding the case-control covariate) Then, after the extracting the fitted factors and loadings that do not correspond to the effects of the covariates, we construct the matrix $\mu \in \mathbb{R}^{p \times n}$, the estimated mean matrix (i.e., exponentiated natural parameter matrix). We then perform a Wilcoxon rank sum test via `stats::wilcox.test` on each gene to perform the DE test, whereby we use a Benjamini-Hochberg multiple-testing correction afterward.

Details of enrichment analysis

For the GO enrichment analysis in Figure 5, we used the `clusterProfiler::enrichGO` function (with `OrgDb = org.Hs.eg.db::org.Hs.eg.db`, `keyType = "SYMBOL"`, `ont = "BP"`, `pAdjustMethod = "BH"`, `pvalueCutoff = 0.05`, and `qvalueCutoff = 0.05`.

For the treemaps as shown in Figure S10, we additionally take the output of `clusterProfiler::enrichGO` and pass it into `rrvgo::calculateSimMatrix` (with `method="Rel"`), `rrvgo::reduceSimMatrix` (setting the `scores` input to be the negative $\log_{10}$ q-values from `clusterProfiler::enrichGO`, and with `threshold=0.7`), and `rrvgo::treemapPlot`.

Additional data analysis results

The following figures provide supplementary results of all these analyses mentioned in the main text.

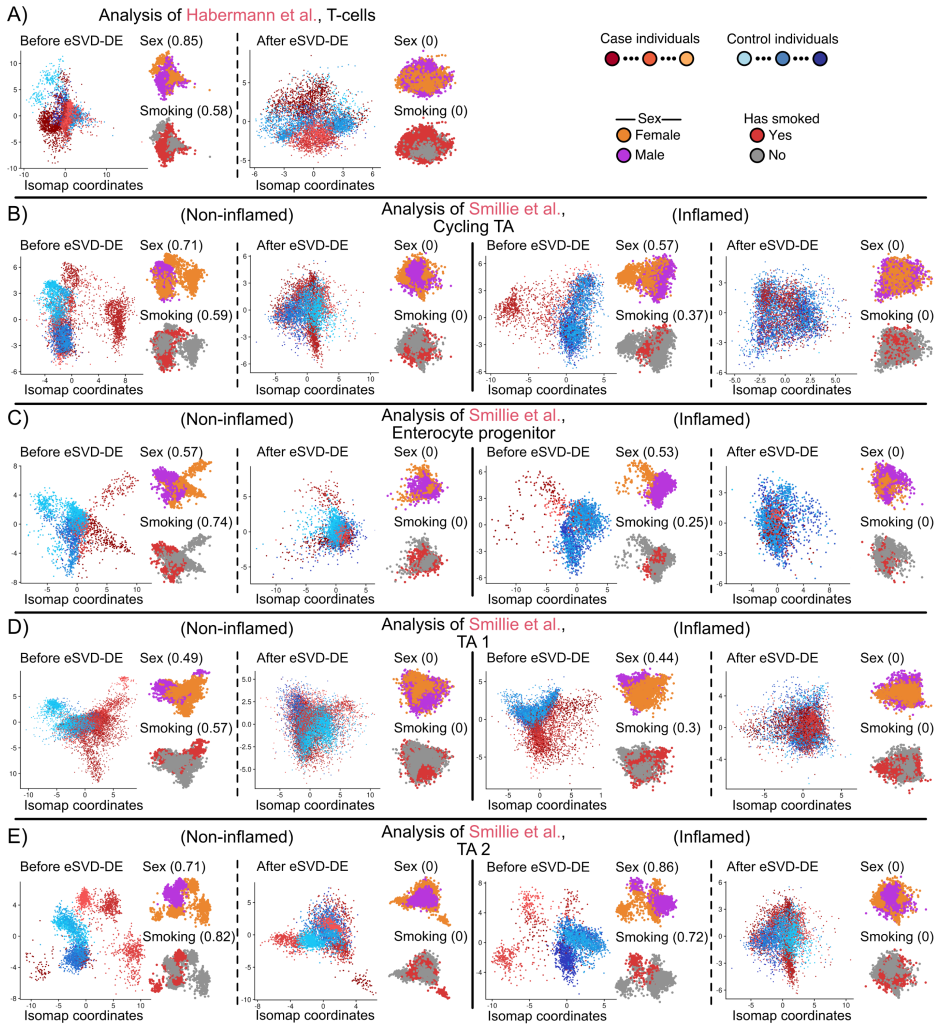

Figure S4 Each Isomap is similar to Figure 3(A,B), but for a different dataset, the Isomap is constructed on the leading principal components (i.e., before eSVD-DE) or based on the low-dimensional cell embedding X after applying the eSVD-DE. We primarily show the Isomaps where the cells are colored by the different case and control individuals that contributed to the cell and also show different insets where the cells are colored by the sex of the individual as well as whether or not the individual smoked. The ideal Isomap after applying eSVD-DE shows no discernable separation between individuals of different sex and smoking status, as these covariates should have been regressed out. **A** Isomaps of the T-cells in the Habermann dataset [19]. **B,C,D,E** Isomaps of the cycling TAs, enterocyte progenitors, TA 1's, or TA 2's in the Habermann dataset [19], either for healthy and non-inflamed cells (left) or healthy and inflamed cells (right).

Supplemental methods

See `https://github.com/linnykos/eSVD2` for the R package `eSVD2` that contains all the functions used in for the eSVD-DE analysis. See `https://github.com/linnykos/eSVD2_examples` for all the scripts used in the analysis.

Consent for publication

Not applicable.

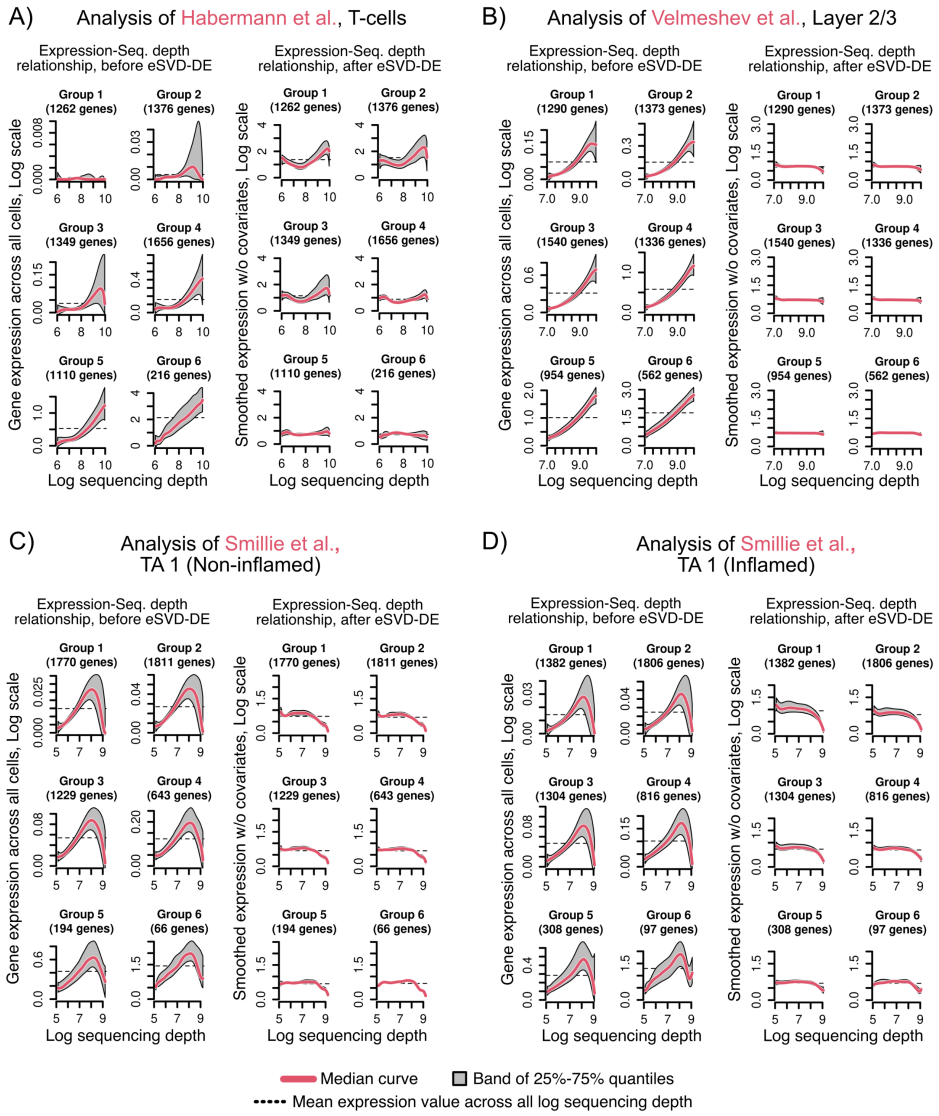

Figure S5 Each line plot is similar to Figure 3(C,D), but for a different dataset, genes are partitioned into six bins based on the gene's mean expression. **A** Relationship between gene expression and sequencing depth for the T-cells in the Habermann dataset [19] before and after applying eSVD-DE. **B** Relationship between gene expression and sequencing depth for the layer 2/3 cells in the Velmeshev dataset [21] before and after applying eSVD-DE. **C,D** Relationship between gene expression and sequencing depth for the TA 1 cells in the Smillie dataset [20] before and after applying eSVD-DE, either for the healthy and inflamed cells (**C**) or the healthy and non-inflamed cells (**D**). An ideal relationship after applying eSVD-DE is that the gene expression is constant for different sequencing depths for each bin of genes.

Availability of data and materials

The Adams data (from [6]) is downloaded from GSE136831, and the individual covariates were provided by the authors of the paper. The Habermann data (from [19]) is downloaded from GSE135893, and the individual covariates were provided by the authors of the paper. The Smilie data (from [20]) is downloaded from https://github.com/cssmillie/ulcerative_colitis, while the individuals' covariates and DE genes are downloaded as Table S1 and S4 from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6662628/. The Velmeshev data (from [21]) is downloaded from https://cells.ucsc.edu/?ds=autism, while the DE genes is downloaded as Data S4 from http://science.sciencemag.org/content/suppl/2019/05/15/364.6441.685.DC1. The housekeeping genes (from [40]) is downloaded as Supplementary Table 1 from https://academic.oup.com/nar/article/49/D1/D947/5871367#supplementary-data. The SFARI genes were downloaded from https://gene.sfari.org/database/gene-scoring/ on January 6, 2022 (using the September

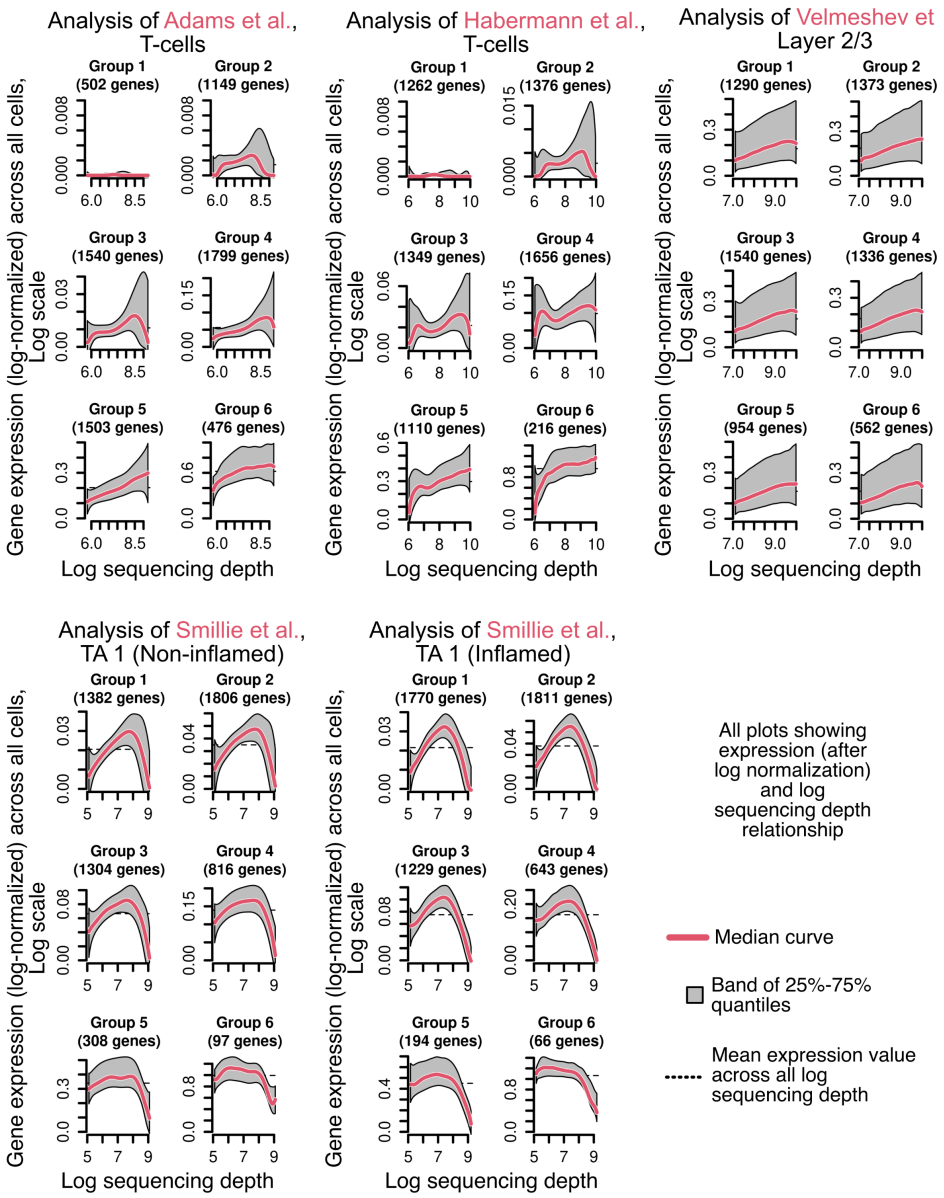

Figure S6 Analogous plots to Figure 3(C,D) and S5, but shown if the gene expressions in each scRNA-seq dataset were log-normalized instead using eSVD-DE. The Y axis denotes the log-normalized gene expression across the same gene partitions as in Figure 3(C,D) and S5 (on an additionally applied log scale). This latter application of the log scale is to keep the plots here as comparable to those in Figure 3(C,D) and S5 as possible. We see that in all the five datasets, log-normalization does not fully mitigate the relationships between gene expression and sequencing depth. This means after log-normalization, genes might still be assess to be differentially expressed solely based on differences in sequencing depth.

2, 2021 release). The Gandal genes (from [49]) is downloaded as Supplementary Data 3 from `https://www.nature.com/articles/s41586-022-05377-7`, where we look at the `DEGene_Statistics` sheet and select the genes with `WholeCortex_ASD_FDR` less than 0.05.

Competing interests

The authors declare that they have no competing interests.

Funding

This project was funded by National Institute of Mental Health (NIMH) grant R01MH123184.

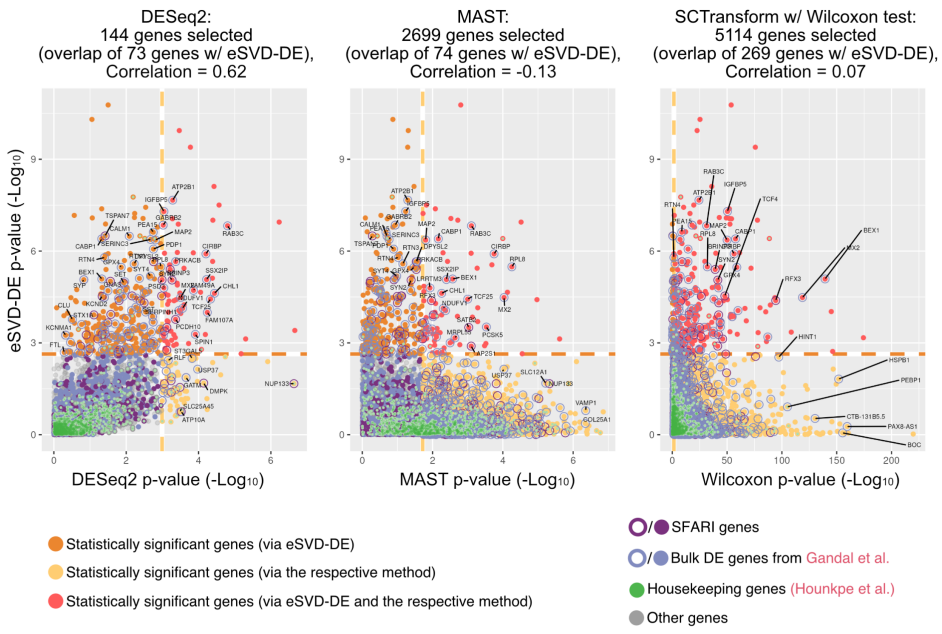

Figure S7 Scatter plots comparing the negative $\log_{10}$ p-values of all 7055 genes derived via eSVD-DE against either DESeq2, MAST, or SCTransform (followed by the Wilcoxon rank-sum test). Genes with a Benjamini-Hochberg multiple-testing corrected p-value less than 0.05 are deemed significant within their respective method. The number of significant genes for each of the three methods is written in the title, alongside the number of genes that overlap with eSVD-DE's 331 significant genes. The correlation between both sets of negative $\log_{10}$ p-values is also noted. Note that both MAST and SCTransform deem a large proportion of the 7055 genes to be statistically significant.

Acknowledgements
We thank Timothy Barry, Jing Lei, Ronald Yurko, and Nancy Zhang for useful comments when developing this method.

Authors' contributions
K.L.Z. and K.R. conceived of the idea. K.L.Z. and Y.Q. coded eSVD-DE in R and C++. K.L.Z. performed the analyses. K.L.Z., Y.Q. and K.R. wrote the paper.

Additional Files
`adams-habermann-T_esvd_results.xlsx` contains the analysis of T-cells of either the Adams data (from [6]) or the Habermann data (from [19]). It contains the estimated log-fold change after removing confounding effects, the eSVD-DE test statistic, the negative $\log_{10}$ p-value, and Benjamini-Hochberg adjusted p-value, and whether the gene is deemed to be a DE gene. Similarly, `smillie_esvd_results.xlsx` contains the same 5 columns, but for cycling TA, enterocyte progenitors, TA 1, and TA 2 when analyzing either inflamed or non-inflamed tissues in the Smillie dataset (from [20]). Lastly, `velmeshev_esvd_results.xlsx` contains the same 5 columns, but for astrocytes, endothelial cells, IN-SST, IN-VIP, layer 2/3, layer 4, layer 5/6, layer 5/6-CC, microglia, oligodendrocytes, and OPC when analyzing the Velmeshev dataset (from [21]).

Author details
[1]Department of Biostatistics, University of Washington, Seattle, USA. [2]School of Statistics & Management, Shanghai University of Finance and Economics, Shanghai, China. [3]Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, USA.

References
1. Kim, N., Kim, H.K., Lee, K., Hong, Y., Cho, J.H., Choi, J.W., Lee, J.-I., Suh, Y.-L., Ku, B.M., Eum, H.H., *et al.*: single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. Nature Communications **11**(1), 1–15 (2020)
2. Gabitto, M., Travaglini, K., Ariza, J., Kaplan, E., Long, B., Rachleff, V., Ding, Y., Mahoney, J., Dee, N., Goldy, J., others, Haynor, D., Gatto, N.M., Jayadev, S., Mutfi, S., Ng, L., Mukherjee, S., Crane, P.K., Latimer, C.S., Levi, B.P., Smith, K., Close, J.L., Miller, J.A., Hodge, R.D., Larson, E.B., Grabowski, T.J., Hawrylycz, M., Keene, C.D., Lein, E.S.: Integrated multimodal cell atlas of alzheimer's disease (2023)
3. Perez, R.K., Gordon, M.G., Subramaniam, M., Kim, M.C., Hartoularos, G.C., Targ, S., Sun, Y., Ogorodnikov, A., Bueno, R., Lu, A., *et al.*: single-cell RNA-seq reveals cell type–specific molecular and genetic associations to lupus. Science **376**(6589), 1970 (2022)

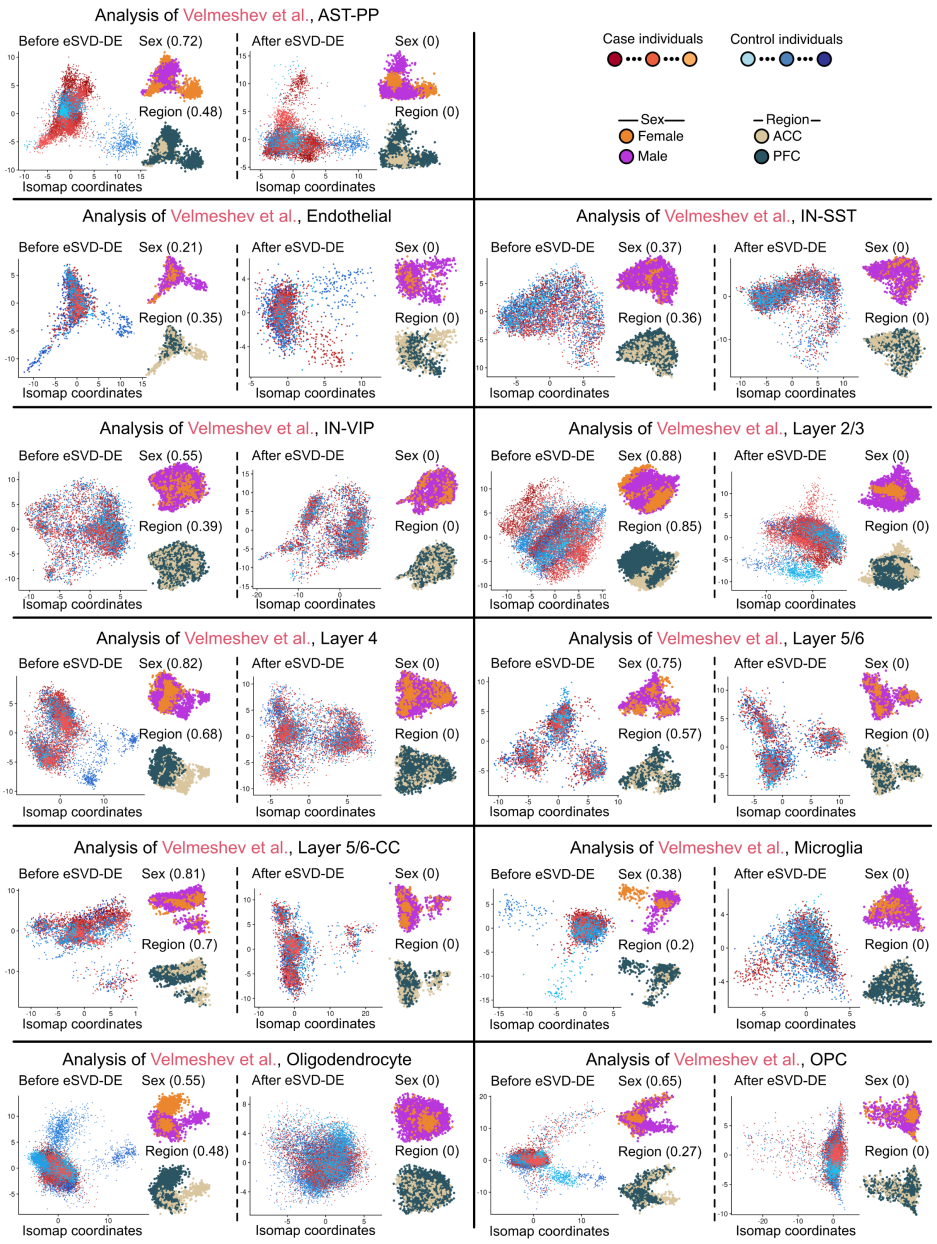

Figure S8 A continuation of Figure S4, where the different rows show the Isomaps for different cell types in the Velmeshev dataset [21], where for any cell type, the left plot shows the Isomap on the leading principal components (i.e., before applying eSVD-DE), and the right plot shows the low-dimensional space X after applying eSVD-DE.

4. Squair, J.W., Gautier, M., Kathe, C., Anderson, M.A., James, N.D., Hutson, T.H., Hudelle, R., Qaiser, T., Matson, K.J., Barraud, Q., Barraud, Q., Levine, A.J., La Manno, G., Skinnider, M.A., Courtine, G.: Confronting false discoveries in single-cell differential expression. Nature Communications **12**(1), 5692 (2021)
5. Mallick, H., Chatterjee, S., Chowdhury, S., Chatterjee, S., Rahnavard, A., Hicks, S.C.: Differential expression of single-cell RNA-seq data using tweedie models. Statistics in Medicine **41**(18), 3492–3510 (2022)
6. Adams, T.S., Schupp, J.C., Poli, S., Ayaub, E.A., Neumark, N., Ahangari, F., Chu, S.G., Raby, B.A., Deluliis, G., Januszyk, M., Duan, Q., Arnett, H.A., Siddiqui, A., Washko, G.R., Homer, R., Yan, X., Rosas, I.O., Kaminski, N.: Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. Science Advances **6**(28), 1983 (2020)
7. Auerbach, B.J., Hu, J., Reilly, M.P., Li, M.: Applications of single-cell genomics and computational strategies to study common disease and population-level variation. Genome Research **31**(10), 1728–1741 (2021)

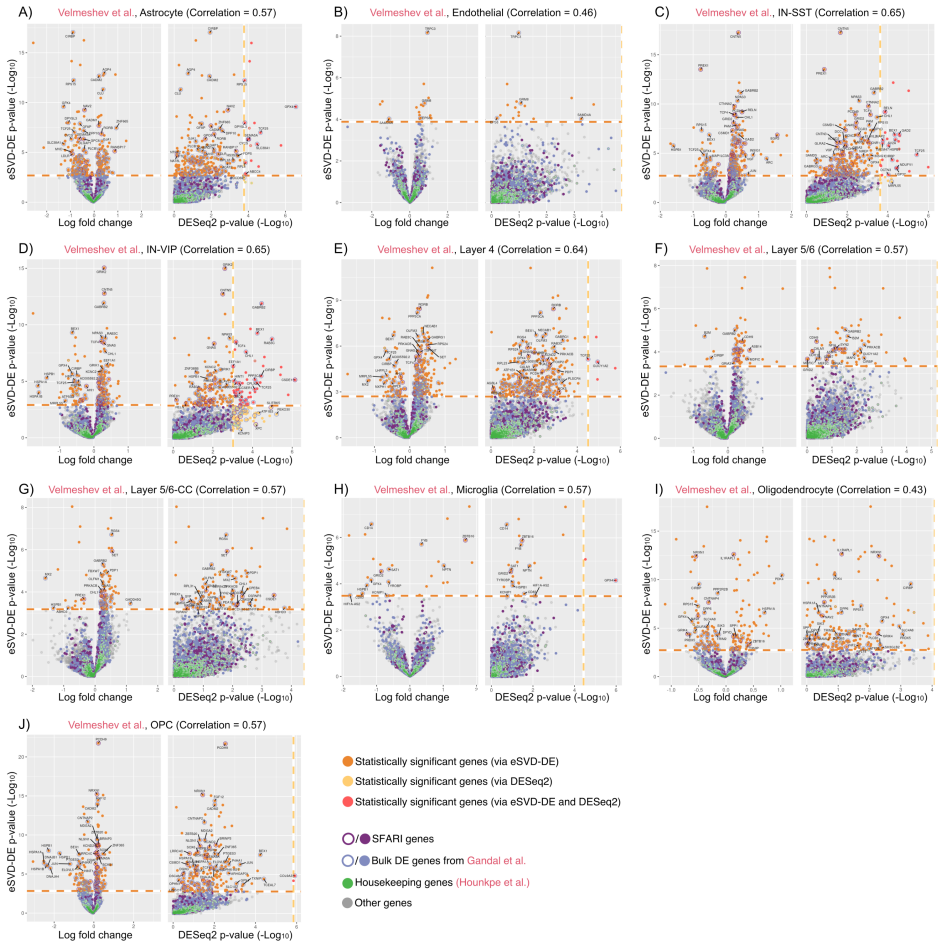

Figure S9 Each volcano plot is similar to Figure 5(B,C), but applied to a different cell type in the Velmeshev dataset [21]. Each gene in every plot is marked differently, depending on whether or not it was deemed significant by eSVD-DE or DESeq2, as well as if the gene was a SFARI gene, a bulk DE gene, or housekeepin gene. The correlation written in the title of each plot denotes the correlation between the negative $\log_{10}$ p-values between the eSVD-DE results and the DESeq2 results. **A** The astrocytes. **B** The endothelial cells. **C** The IN-SST cells. **D** The IN-VIP cells. **E** The layer 4 cells. **F** The layer 5/6 cells. **G** The layer 5/6-cc cells. **H** The microglias **I** The oligodendrocytes. **J** The OPCs.

8. Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology **15**(12), 550 (2014)
9. McCarthy, D.J., Chen, Y., Smyth, G.K.: Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. Nucleic Acids Research **40**(10), 4288–4297 (2012)
10. Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., Slichter, C.K., Miller, H.W., McElrath, M.J., Prlic, M., Linsley, P.S., Gottardo, R.: MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biology **16**(1), 278 (2015)
11. He, L., Davila-Velderrain, J., Sumida, T.S., Hafler, D.A., Kellis, M., Kulminski, A.M.: NEBULA is a fast negative binomial mixed model for differential or co-expression analysis of large-scale multi-subject single-cell data. Communications biology **4**(1), 629 (2021)
12. 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 Biology **20**(1), 1–16 (2019)
13. 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. Nature Communications **9**(1), 284 (2018)
14. Nicol, P.B., Miller, J.W.: Model-based dimensionality reduction for single-cell RNA-seq using generalized bilinear models. bioRxiv, 2023–04 (2023)
15. Andrews, T.S., Hemberg, M.: False signals induced by single-cell imputation. F1000Research **7** (2018)
16. Boyeau, P., Regier, J., Gayoso, A., Jordan, M.I., Lopez, R., Yosef, N.: An empirical bayes method for

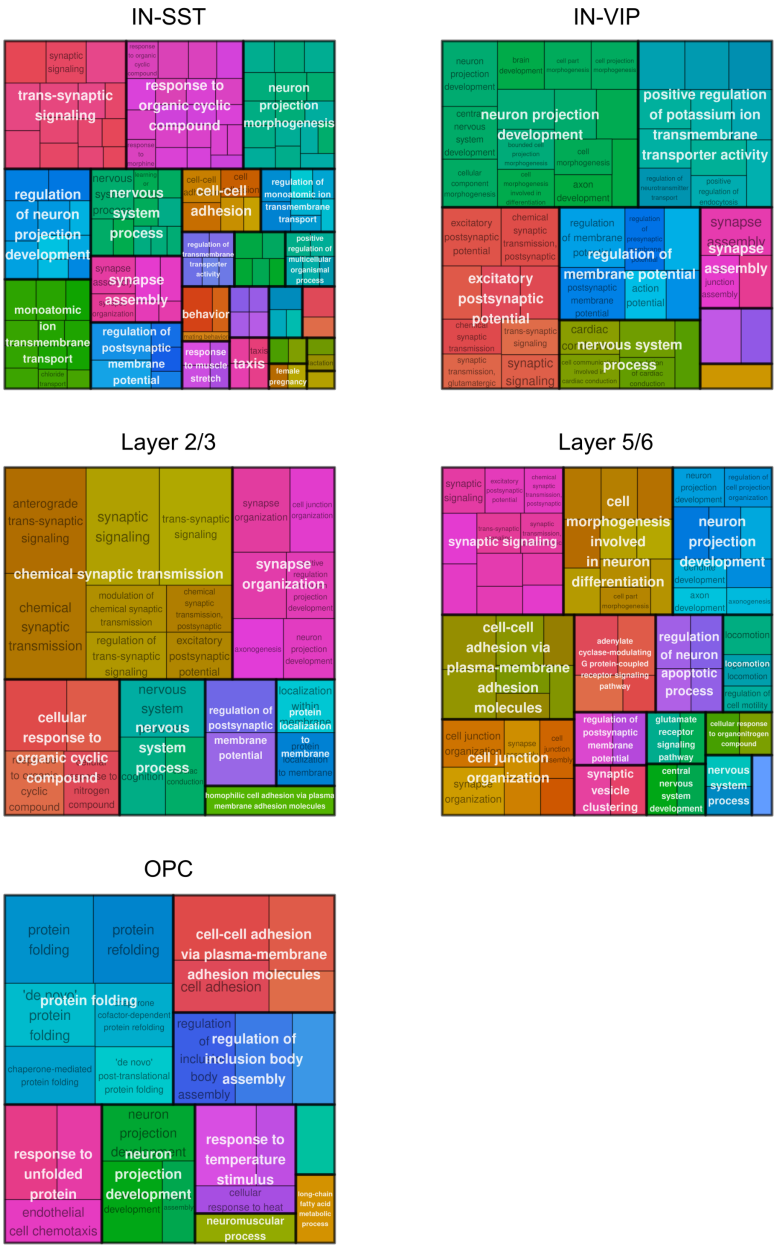

Figure S10 Tree maps of the enriched GO terms among the significant genes discovered by eSVD-DE when analyzing different cell types. While this analysis was run on every cell type, some significant genes were not enriched enough GO terms to form a tree map (and hence, not shown among the five shown in this plot). Each plot shows different enriched GO terms (deemed significant by Fisher's exact test p-value after Benjamini-Hochberg multiple-testing correction), grouped by two hierarchies, where the top hierarchy is shaded with a different color.

differential expression analysis of single cells with deep generative models. Proceedings of the National Academy of Sciences **120**(21), 2209124120 (2023)

17. Weinberger, E., Lin, C., Lee, S.-I.: Isolating salient variations of interest in single-cell data with contrastiveVI. Nature Methods, 1–10 (2023)
18. Lin, K.Z., Lei, J., Roeder, K.: Exponential-family embedding with application to cell developmental trajectories for single-cell RNA-seq data. Journal of the American Statistical Association **116**(534), 457–470 (2021)
19. Habermann, A.C., Gutierrez, A.J., Bui, L.T., Yahn, S.L., Winters, N.I., Calvi, C.L., Peter, L., Chung, M.-I., Taylor, C.J., Jetter, C., Raju, L., Roberson, J., Ding, G., Wood, L., Sucre, J.M.S., Richmond, B.W., Serezani,

A.P., McDonnell, W.J., Mallal, S.B., Bacchetta, M.J., Loyd, J.E., Shaver, C.M., Ware, L.B., Bremner, R., Walia, R., Blackwell, T.S., Banovich, N.E., Kropski, J.A.: Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. Science Advances **6**(28), 1972 (2020)

20. Smillie, C.S., Biton, M., Ordovas-Montanes, J., Sullivan, K.M., Burgin, G., Graham, D.B., Herbst, R.H., Rogel, N., Slyper, M., Waldman, J., Sud, M., Andrews, E., Velonias, G., Haber, A.L., Jagadeesh, K., Vickovic, S., Yao, J., Stevens, C., Dionne, D., Nguyen, L.T., Villani, A.-C., Hofree, M., Creasey, E.A., Huang, H., Rozenblatt-Rosen, O., Garber, J.J., Khalili, H., Desch, A.N., Daly, M.J., Ananthakrishnan, A.N., Shalek, A.K., Xavier, R.J., Regev, A.: Intra-and inter-cellular rewiring of the human colon during ulcerative colitis. Cell **178**(3), 714–730 (2019)
21. Velmeshev, D., Schirmer, L., Jung, D., Haeussler, M., Perez, Y., Mayer, S., Bhaduri, A., Goyal, N., Rowitch, D.H., Kriegstein, A.R.: Single-cell genomics identifies cell type–specific molecular changes in autism. Science **364**(6441), 685–689 (2019)
22. Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J.I., Raj, A., Li, M., Zhang, N.R.: SAVER: Gene expression recovery for single-cell RNA sequencing. Nature Methods **15**(7), 539–542 (2018)
23. Gayoso, A., Steier, Z., Lopez, R., Regier, J., Nazor, K.L., Streets, A., Yosef, N.: Joint probabilistic modeling of single-cell multi-omic data with totalVI. Nature Methods **18**(3), 272–282 (2021)
24. van Iterson, M., van Zwet, E.W., Heijmans, B.T.: Controlling bias and inflation in epigenome-and transcriptome-wide association studies using the empirical null distribution. Genome biology **18**(1), 1–13 (2017)
25. Sarkar, A., Stephens, M.: Separating measurement and expression models clarifies confusion in single cell RNA-seq analysis. Nature Genetics **53**(6), 770–777 (2021)
26. Chen, W., Li, Y., Easton, J., Finkelstein, D., Wu, G., Chen, X.: UMI-count modeling and differential expression analysis for single-cell RNA sequencing. Genome Biology **19**(1), 70 (2018)
27. Zhang, M., Liu, S., Miao, Z., Han, F., Gottardo, R., Sun, W.: IDEAS: Individual level differential expression analysis for single-cell RNA-seq data. Genome Biology **23**(1), 1–17 (2022)
28. Korthauer, K.D., Chu, L.-F., Newton, M.A., Li, Y., Thomson, J., Stewart, R., Kendziorski, C.: A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. Genome Biology **17**(1), 1–15 (2016)
29. Schefzik, R., Flesch, J., Goncalves, A.: Fast identification of differential distributions in single-cell RNA-sequencing data with waddR. Bioinformatics **37**(19), 3204–3211 (2021)
30. Liu, Y., Zhao, J., Adams, T.S., Wang, N., Schupp, J.C., Wu, W., McDonough, J.E., Chupp, G.L., Kaminski, N., Wang, Z., Yan, X.: iDESC: Identifying differential expression in single-cell RNA sequencing data with multiple subjects. BMC Bioinformatics **24**(1), 318 (2023)
31. Junttila, S., Smolander, J., Elo, L.L.: Benchmarking methods for detecting differential states between conditions from multi-subject single-cell RNA-seq data. Briefings in Bioinformatics **23**(5), 286 (2022)
32. Tenenbaum, J.B., Silva, V.d., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290**(5500), 2319–2323 (2000)
33. Kraemer, G., Reichstein, M., Mahecha, M.D.: dimRed and coRanking - unifying dimensionality reduction in R. R Journal **10**(1), 342–358 (2018)
34. McInnes, L., Healy, J., Melville, J.: UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)
35. Ham, J., Lee, D.D., Mika, S., Schölkopf, B.: A kernel view of the dimensionality reduction of manifolds. In: Proceedings of the Twenty-first International Conference on Machine Learning, p. 47 (2004)
36. Wu, H.-T., Wu, N., *et al.*: Think globally, fit locally under the manifold setup: Asymptotic analysis of locally linear embedding. The Annals of Statistics **46**(6B), 3805–3837 (2018)
37. Perturbation bounds for procrustes, classical scaling, and trilateration, with applications to manifold learning. Journal of machine learning research **21** (2020)
38. Li, Y., Ge, X., Peng, F., Li, W., Li, J.J.: Exaggerated false positives by popular differential expression methods when analyzing human population samples. Genome Biology **23**(1), 1–13 (2022)
39. Hafemeister, C., Satija, R.: Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biology **20**(1), 1–15 (2019)
40. Hounkpe, B.W., Chenou, F., de Lima, F., De Paula, E.V.: HRT atlas v1.0 database: Redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. Nucleic Acids Research **49**(D1), 947–955 (2021)
41. Risso, D., Ngai, J., Speed, T.P., Dudoit, S.: Normalization of RNA-seq data using factor analysis of control genes or samples. Nature Biotechnology **32**(9), 896–902 (2014)
42. Lause, J., Berens, P., Kobak, D.: Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. Genome Biology **22**(1), 1–20 (2021)
43. Cole, M.B., Risso, D., Wagner, A., DeTomaso, D., Ngai, J., Purdom, E., Dudoit, S., Yosef, N.: Performance assessment and selection of normalization procedures for single-cell RNA-seq. Cell systems **8**(4), 315–328 (2019)
44. Lun, A.T., Bach, K., Marioni, J.C.: Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biology **17**(1), 75 (2016)
45. Nance, T., Smith, K.S., Anaya, V., Richardson, R., Ho, L., Pala, M., Mostafavi, S., Battle, A., Feghali-Bostwick, C., Rosen, G., Montgomery, S.B.: Transcriptome analysis reveals differential splicing events in IPF lung tissue. PloS One **9**(3), 92111 (2014)
46. Joshi, N., Watanabe, S., Verma, R., Jablonski, R.P., Chen, C.-I., Cheresh, P., Markov, N.S., Reyfman, P.A., McQuattie-Pimentel, A.C., Sichizya, L., Lu, Z., Piseaux, R., Kirchenbuechler, D., Flozak, A.S., Gottardi, C.J., Cuda, C.M., Perlman, H., Jain, M., Kamp, D.W., Budinger, G.R.S., Misharin, A.V.: A spatially restricted fibrotic niche in pulmonary fibrosis is sustained by M-CSF/M-CSFR signalling in monocyte-derived alveolar macrophages. European Respiratory Journal **55**(1) (2020)
47. Gauldie, J., Kolb, M., Ask, K., Martin, G., Bonniaud, P., Warburton, D.: Smad3 signaling involved in pulmonary fibrosis and emphysema. Proceedings of the American Thoracic Society **3**(8), 696–702 (2006)

48. SFARI gene database (2022). `https://gene.sfari.org/` Accessed 2022-10-20
49. Gandal, M.J., Haney, J.R., Wamsley, B., Yap, C.X., Parhami, S., Emani, P.S., Chang, N., Chen, G.T., Hoftman, G.D., de Alba, D., Ramaswami, G., Hartl, C.L., Bhattacharya, A., Luo, C., Jin, T., Wang, D., Kawaguchi, R., Quintero, D., Ou, J., Wu, Y.E., Parikshak, N.N., Swarup, V., Belgard, T.G., Gerstein, M., Pasaniuc, B., Geschwind, D.H.: Broad transcriptomic dysregulation occurs across the cerebral cortex in ASD. Nature **611**(7936), 532–539 (2022). doi:10.1038/s41586-022-05377-7
50. Efron, B.: Microarrays, empirical Bayes and the two-groups model. Statistical Science **23**(1), 1–22 (2008)
51. Read, D.F., Daza, R.M., Booth, G.T., Jackson, D.L., Gladden, R.G., Srivatsan, S.R., Ewing, B., Franks, J.M., Spurrell, C.H., Gomes, A.R., O'Day, D., Gogate, A.A., Martin, B.K., Starita, L., Lin, Y., Shendure, J., Lin, S., Trapnell, C.: Single-cell analysis of chromatin and expression reveals age-and sex-associated alterations in the human heart. bioRxiv (2022)
52. Agarwal, D., Wang, J., Zhang, N.R., *et al.*: Data denoising and post-denoising corrections in single cell RNA sequencing. Statistical Science **35**(1), 112–128 (2020)
53. Wang, L., Zhang, X., Gu, Q.: A unified computational and statistical framework for nonconvex low-rank matrix estimation. Artificial Intelligence and Statistics, 981–990 (2017). PMLR
54. Li, X., Lu, J., Arora, R., Haupt, J., Liu, H., Wang, Z., Zhao, T.: Symmetry, saddle points, and global optimization landscape of nonconvex matrix factorization. IEEE Transactions on Information Theory **65**(6), 3489–3514 (2019)
55. Efron, B.: Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. Journal of the American Statistical Association **99**(465), 96–104 (2004)
56. Ma, Y., Sun, S., Shang, X., Keller, E.T., Chen, M., Zhou, X.: Integrative differential expression and gene set enrichment analysis using summary statistics for scRNA-seq studies. Nature Communications **11**(1), 1–13 (2020)
57. Gao, Q., Ji, Z., Wang, L., Owzar, K., Li, Q.-J., Chan, C., Xie, J.: SifiNet: A robust and accurate method to identify feature gene sets and annotate cells. bioRxiv, 2023–05 (2023)