METHOD

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

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

[*]Correspondence:
kevinl1@wharton.upenn.edu
[1]Department of Statistics & Data Science, University of Pennsylvania, Philadelphia, 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 clinician studies, but lack methods designed to investigate differentially-expressed (DE) genes among individuals. While there exist numerous methods to find DE genes using bulk sequencing data or scRNA data from limited individuals, differential-expression tests among large cohorts of case and control individuals using scRNA data poses unique challenges due to substantial effects of cohort-level confounding covariates that are difficult to account for in the presence of sparsely-observed genes. In this paper, we develop a differential-expression test based on the exponential-family SVD (eSVD) which pools information across genes and removes confounding covariate effects via a matrix factorization approach. We then test for significant mean differences between the case and control individuals' posterior mean distributions. The matrix factorization approach enables abundant diagnostic and tuning strategies. We demonstrate the accuracy and power-increase of our method when compared to other DE methods typically used to repurposed for analyzing cohort-level differential expression based on simulated data, and different datasets studying individuals with idiopathic pulmonary fibrosis, varying severity of ulcerative colitis, and autism spectrum disorder, all when compared to their respective control individuals.

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

Background

High-throughput single-cell RNA-seq (scRNA) has advanced tremendously over the last decade, both in technology and in methods, 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, an insight that was previously inaccessible with bulk RNA-seq technology. As the technology has developed more in its 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 population. For example, biologists have sequenced hundreds of thousands of cells for 44 individuals with lung adenocarcinoma [1] as well as millions of blood cells for 162 individuals with systemic lupus erythematosus and 99 control individuals [2] in order to discover DE genes for the respective disease/disorder consistent with the entire cohort. However, despite the hundreds of existing DE methods designed for

single-cell data that have been thoroughly benchmarked in recent years [3, 4], only a handful of methods are designed for such multi-individual scRNA datasets. This is because most existing DE methods are designed to find differentially-expressed patterns among *cells*, whereas these cohort-wide studies are instead interested in differentially-expressed patterns among *individuals*. This pressing methodological gap has been raised by many biologists collecting these multi-individual scRNA datasets [5, 6]. Therefore, we combine the ideas from DE methods previously made for cohort-wide bulk RNA-seq studies along with modern DE methods designed for scRNA data to design a cohort-level DE method for scRNA data. We note that while recent technological developments in spatial transcriptomics and jointly-sequenced multiomics platforms have started to be sequenced at the single-cell level for cohorts, the statistical models and understandings of scRNA data is much more mature. This enables us to study more principled ways to model multi-individual scRNA data, which we hope could be used to inspire multi-individual DE methods for other single-cell technologies in the future, as well as pioneer applications in future eQTL and *in vivo* CRISPR analyses.

One of the main distinctions between DE analyses among cells compared to among individuals lies in how the difference between the cases' and controls' population models is quantified. DE analyses among individuals need to account for both variability within an individual as well as variability among individuals, and most existing DE methods lack one of these two aspects. On one hand, variability within an individual hinders "pseudo-bulk" analyses where all the cells among each individual summed to yield a bulk sample. This is because even if the case and control individuals have well-separated mean expressions for a particular gene, this difference might be statistically insignificant when accounting for the variability of expression within each individual. This variability is lost when constructing the pseudo-bulk samples. On the other hand, variability across individuals hinders most DE methods made for scRNA. This is because if individuals contribute disproportionate amount of cells of a particular cell type to the study and the variability among individuals is large, the mean difference between case and control *cells* could be significant even if the mean difference between *individuals* is insignificant.

The second main distinction is that in multi-individual scRNA data, there are potentially dominant effects of cohort-level covariates such as age, gender, 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. To address this, a conventional strategy for bulk RNA-seq data is to remove the confounding covariate effect gene-by-gene via a generalized linear model (GLM), such as in DESeq2 [7]. This strategy can be much less effective for multi-individual scRNA data though since each gene is sequenced much more sparsely by comparison. This is in contrast with scRNA data for cell-lines or clonal mice, where the effects of confounding covariates is more mild [8]. Hence, a dimension-reduction method such as GLM-PCA [9] and ZINB-WaVE [10]. is ideal, as information across genes is pooled together to aid in effectively removing the effects of the confounding covariates. However, a naive application of DE on the dimension-reduced scRNA has been observed to inflate the Type-I error [11]. This is because highly DE genes can have expression patterns not sufficiently captured by a case-control indicator covariate, and the residual expression's high variability can distort the low-dimension embedding.

To overcome these two main obstacles, we design the exponential-family SVD (eSVD), a dimension-reduction method that extends of our previous work [12], but is now computationally improved and also coupled with an appropriate downstream workflow to account for all the aforementioned obstacles. Importantly, our statistical model assess the posterior mean and posterior variance of the gene expression, which helps counteract the inflation to Type I error. This posterior idea has its roots in SAVER [13], but is adapted for the setting for assessing DE for multi-individual studies. This combination of dimension reduction and the posterior distribution allows more effective pooling of information across genes. Hence, eSVD better removes the confounding covariates' effects compared to existing strategies to detect DE genes in multi-individual studies and also enables model diagnostics to assess if the assumed statistical model is appropriate modeling the scRNA dataset.

In this paper, we focus on the setting where the cells in a UMI-based scRNA dataset have been labeled by their cell-types, and we are interested in the DE of mean expressions among one or a few of these cell-types. We show that eSVD can find reproducible signals in multiple pairs of multi-individual datasets, either across various cell-types between two independent studies of idiopathic pulmonary fibrosis (IPF) in the human lung [5, 14] or within a study of non-inflamed and inflamed cells studying ulcerative colitis in the human colon [15]. We also show that eSVD can find novel DE genes across different cell-types in a dataset studying autism [16]. Altogether, these analysis are intended to provide evidence that eSVD is a useful tool for future biologists to investigate differential expression among multi-individual studies that will become more prevalent as high-throughput sequencing technologies are applied to large cohorts.

Results

Overview of eSVD for differential expression testing

eSVD 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 [9] and ZINB-WaVE [10], where we embed the scRNA dataset via the Poisson distribution alongside covariate information (Figure 1A). Importantly, these covariates contain an intercept term, the log-library size (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) or cell-level covariates such as percent mitochondrial. Towards this end, eSVD learns a coefficient matrix that removes the effects of the covariates as well as low-dimensional embedding of "residuals." While this low-dimensional embedding can be useful for a wide variety of applications, we focus on its application to DE analyses for multi-individual data in this paper.

After fitting a low-dimensional embedding, we cater the following procedure to test for DE among individuals, an aspect absent from our previous work [12]. First, 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 is computed by first assessing the dispersion of each gene, which measures how much extra variability exists between the assumed Poisson fit in the

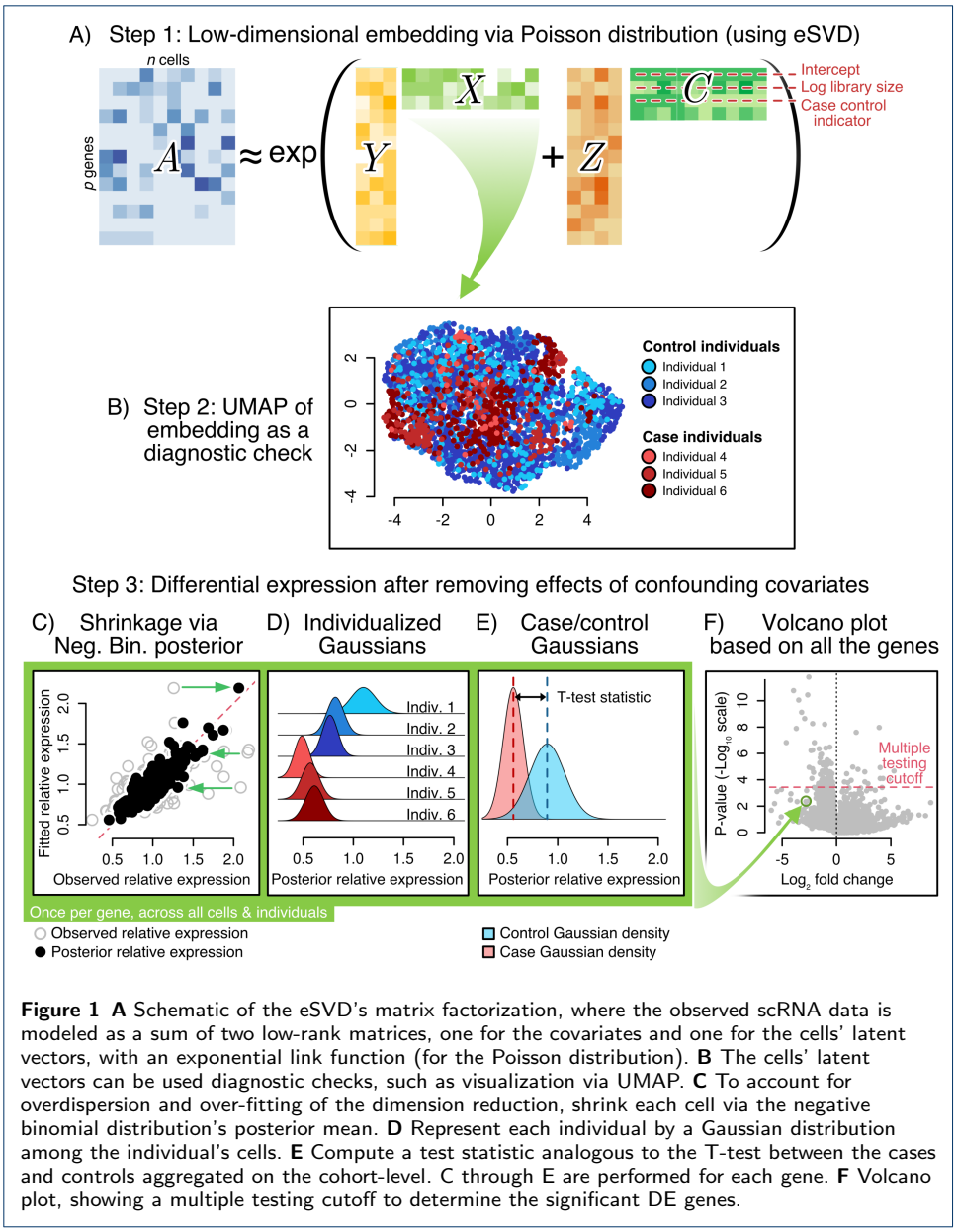

Figure 1 **A** Schematic of the eSVD's matrix factorization, where the observed scRNA data is modeled as a sum of two low-rank matrices, one for the covariates and one for the cells' latent vectors, with an exponential link function (for the Poisson distribution). **B** The cells' latent vectors can be used diagnostic checks, such as visualization via UMAP. **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 between the cases and controls aggregated on the cohort-level. C through E are performed for each gene. **F** Volcano plot, showing a multiple testing cutoff to determine the significant DE genes.

low-dimensional embedding and the observed data, analogous to works like SAVER [13] and totalVI [17]. 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 in accordance with 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 [18] (Figure 1F). More details of the eSVD procedure are described in the Appendix.

eSVD relies on three primary statistical assumptions: (1) scRNA data can be appropriately modeled through the Gamma-Poisson distribution, (2) the effects of confounding covariates can be effectively be removed through a GLM framework, and (3) the DE genes show significant difference in means between case and control individuals, after accounting for the cohort-level covariates. Towards the first assumption, our posterior distribution effectively models counts through a Negative Binomial distribution, which has been justified for modeling scRNA data [19] and has served as the foundation for many methods [10, 20, 13]. Towards the second assumption, many methods from the bulk RNA-seq data such as edgeR [21] and DESeq2 [7] 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 [22], scDD [23] and waddR [24] that test for differential *distributions* instead of differential mean expressions. However, we have have found it harder to generalize these results when comparing across different datasets of similar biological systems, as differentially-distributed genes are not neatly characterized by over-/under-expression.

Design of simulation studies

We design an overall simulation suite to illustrate two aspects of DE for multi-individual scRNA-seq: 1) testing for differential expression among individuals is fundamentally different from among cells and 2) eSVD's framework enables more accurate inference due to its usage of dimension reduction and posterior correction. The former goal can be conceptualized through Figure 2A. Large within-individual variability hinders "pseudo-bulk" 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. Large between-individual variability hinders most existing single-cell methods designed for testing DE among cells, especially if different individuals contribute drastically different proportions of cells. This is because even if the difference in mean expression is insignificant on the cohort-level, a large number of cells from a few case and control individuals can yield highly significant differences in mean expression on the cell-level. Since these distinctions have been explained in previous work [22, 25, 26], we defer simulations that highlight this former goal to the Appendix. Instead, we focus presenting results to illustrate the latter goal here.

The simulation study contains 700 genes and 20 individuals equally split among cases and controls, where cells contributed by each individual has gene expression profiles that are impacted by covariates correlated with the case-control status (such as age, sex and tobacco-use), and the gene expressions sampled from a Negative Binomial distribution (Figure 2B). Here, we focus on the setting where each individual contributes 250 cells, and the true generative model is more complex than the statistical model assumed by eSVD in order to not give eSVD an unfair advantage. Specifically, eSVD's assumed statistical model of both the relation between genes and confounding variables as well as the low-dimensional manifold that the cells lie on are misspecified with respect to the true generative model. Nonetheless, the data is generated such that a set of 50 genes are drastically more differentially expressed among cases and controls on the cohort-level compared to the remaining 650 genes,

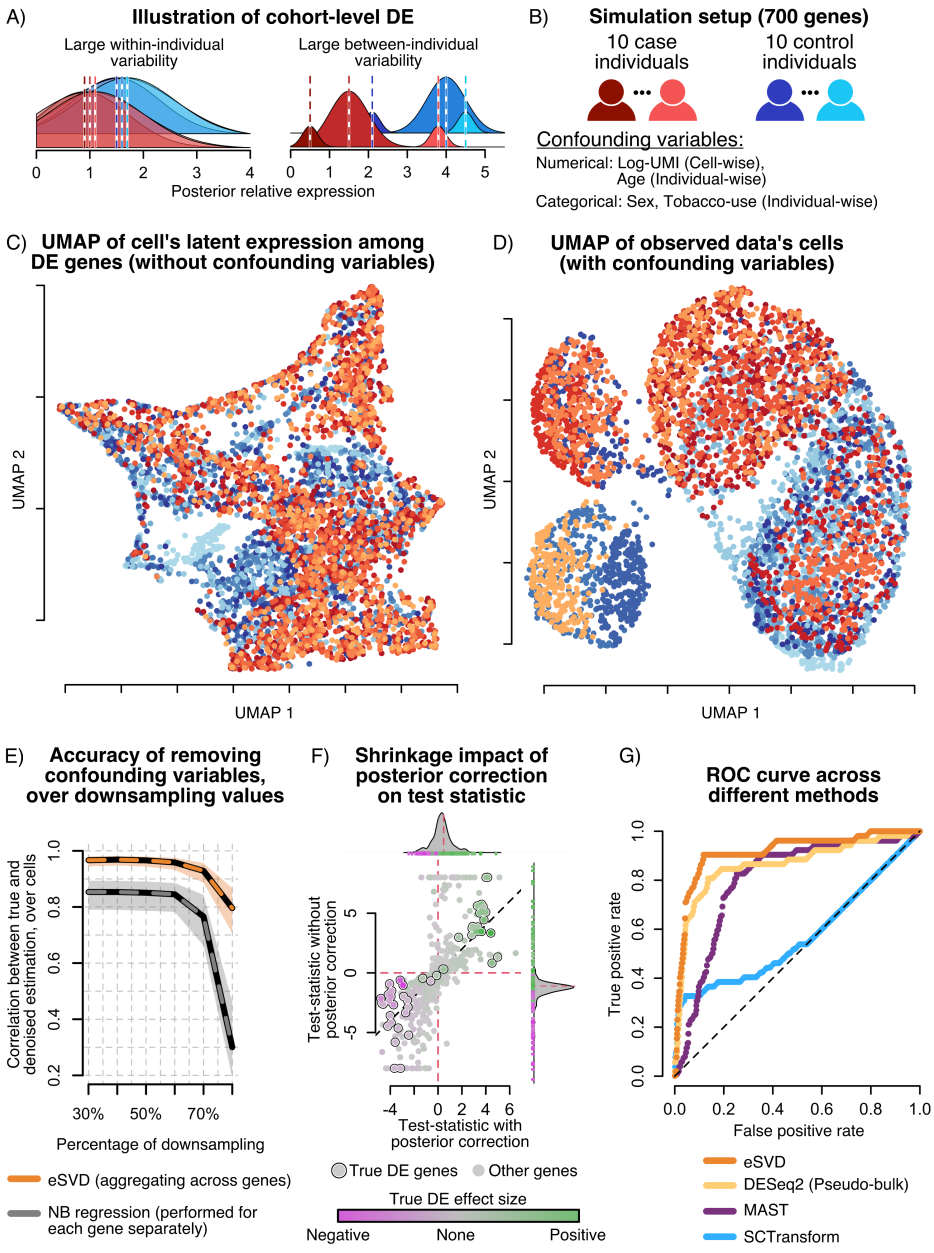

Figure 2 **A** Illustration of challenges for cohort-level DE testing. **B** Setup for our simulation setup. **C** UMAP of the cells' latent vectors, not impacted any confounding variables. We see that no individuals' concentrate tightly in any region on the UMAP manifold, and the mild separation between the cases (shades of red) and controls (shades of blue) suggest some DE genes. **D** UMAP of the observed data. We see large differences between cases and controls induced by confounding variables. **E** Downsampling experiment, demonstrating that by pooling information across genes, eSVD 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 represents each gene's test statistic with and without posterior correction respectively. The genes are colored by their true DE effect size, of which the circled genes denote the "true" DE genes. **G** ROC curve comparing four different methods, illustrating that eSVD has more power than competing methods.

barring confounding effects. This true signal can be visualized as a UMAP, where no apparent stratification is observed within the case individuals or within the control

individuals, but the signal of DE genes separate the cases from controls (Figure 2C). However, the methods in our simulation study are applied to data where covariates such as age, sex and tobacco-use confound which genes are DE (Figure 2D).

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

We first show that by pooling information across genes via the dimension-reduction framework, eSVD removes the confounding variables' effects more accurately compared to a Negative Binomial (NB) regression applied separately for each gene (Figure 2E). To demonstrate this, after generating the synthetic data, we incrementally downsample the data to make the differentially-expressed signals more feint. As the amount of downsampling increases, the more difficult it is for a method to properly remove the confounding effects. 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 removes the confounding effects more accurately than the gene-by-gene NB regression (0.95 to 0.87). Additionally, even though both methods drop in accuracy for downsampling levels of 60% or more, eSVD still is 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 that enables an accurate estimation of the NB regression coefficients. However, by pooling information across genes, the coefficients for the confounding variables are better estimated.

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

We next illustrate that the posterior-correction performed by eSVD is pivotal for DE testing on the cohort-level through our simulation (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 "true" DE genes. (Recall, we purposely design our model to be misspecified in order to not give any particular method an advantage. However, this means more care is required when defining what the "true" DE genes are in such settings. See Appendix for more details.) 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 observe an inflation of Type-I error for such procedures [26, 27, 11]. However, by adjusting the relative expressions using the posterior mean, many of the null genes have drastically smaller test-statistic (y-axis). This phenomenon occurs because the assumed linear model between the covariates and gene's natural parameter does not sufficiently capture more complex non-parametric relations often displayed among individuals, and hence distort the lower-dimensional embedding of the "residuals." This has an adverse effect when genes varying substantially in sparsity – genes that are sparsely observed substantially denoised by projecting the cells against the incorrect manifold.

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

Lastly, we lastly illustrate the eSVD has more power than other conventional methods to test for DE in multi-individual scRNA data through our simulation. Since different methods estimate sets of DE genes with dramatically different sizes if given 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 [7] (i.e., the prototypical method representing "pseudo-bulk" analyses), MAST [28] (i.e., the commonly used method using mixed-effect models where there is a random effect among individuals) and SCTransform [29] (i.e., the prototypical method of performing DE ignoring the individual structure). We observe that eSVD has the highest power when compared 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 next best since it accounts for the individual structure.

eSVD enables diagnostics to assess the performance of removing covariate effects

Moving beyond simulations, we now investigate how well eSVD removes confounding effects in scRNA datasets and how the dimension-reduction framework enables conventional model diagnostics. To demonstrate this, we investigate a broad collection of scRNA 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 [5]. 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 [30]. When visualizing these cells via UMAP [31], 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 cohort-level covariates like sex and smoking status also are locally concentrated in different regions of the UMAP. In both of these instances, we do not wish to report genes as differentially-expressed simply because they are indicative of sex differences or damage induced by smoking.

After applying eSVD, we see that resulting UMAP of the fitted embedding no longer carries expression patterns that are correlated with the individuals, sex or smoking status (Figure 3B). In fact, UMAP such as these can help biologists assess (1) how well the assumed statistical model approximates the scRNA data and (2) what the relative effects each covariate bears on the gene expression. Towards the first point, if biologists prefer visual diagnostics, they can plot the UMAP of the fitted data with all the covariate effects included to assess if latent dimensionality of the embedding is sufficient. If biologists prefer quantitative diagnostics, the eSVD framework can used on scRNA datasets with missing values. This means we can purposely omit a small percentage of values from the scRNA count matrix prior to applying eSVD, and assess how correlated the predicted values are to these omitted values. This was used successfully in our previous work [12]. Towards the second point, biologists can plot the UMAP with any covariate effects (such as those in Figure 3B), then sequentially make additional UMAPs that include the

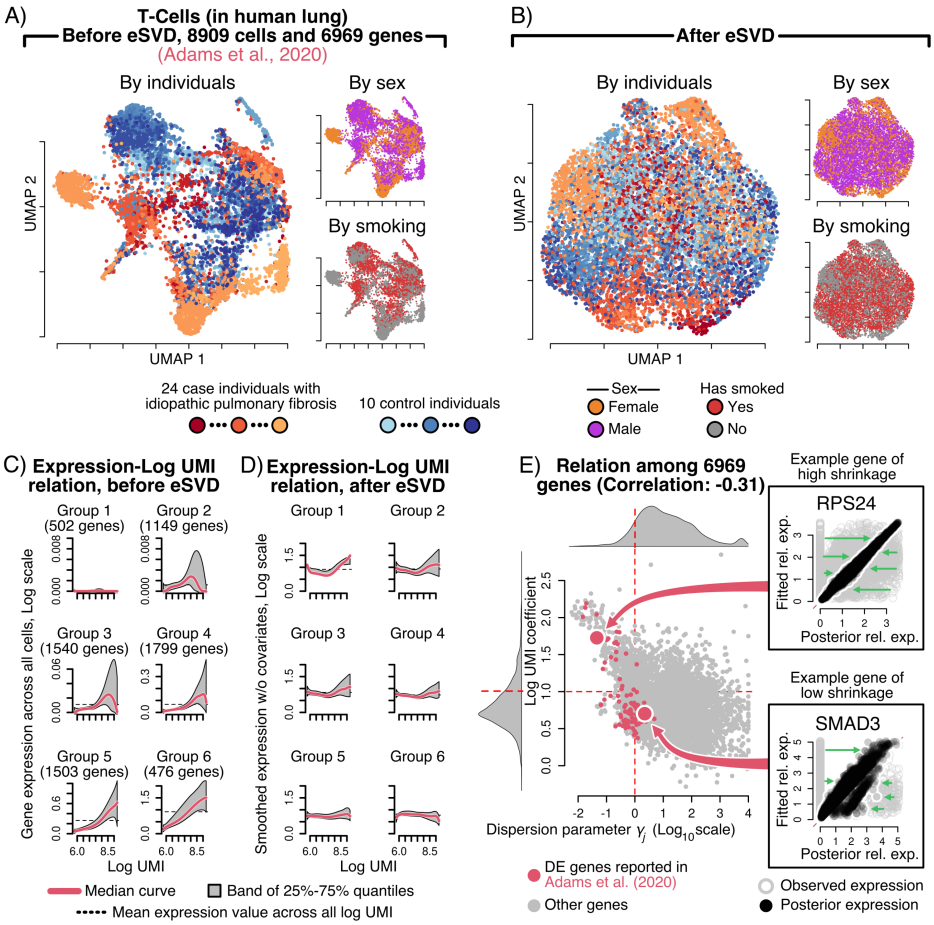

Figure 3 **A** UMAP of T-cells in the Adams dataset, shown by individuals, sex and smoking status. **B** UMAP after eSVD, demonstrating that confounding covariate and individual effects have been removed. **C,D** Relation between gene expression and cells' log UMIs prior to or after eSVD respectively, for groups of 6 genes partitioned by the gene's mean expression. relation after eSVD. After eSVD, each gene becomes more uncorrelated with the cell's log UMI. **E** Relation between each cell's Log UMI and dispersion parameter, highlighting two genes displaying different levels of shrinkage via the posterior mean.

effect of one more covariate sequentially (see Appendix). In this way, biologists can visually assess with of the confounding covariates had the most profound impact on the underlying gene expression. For all these reasons, we advocate the matrix-factorization strategy in the eSVD as opposed to deep-learning alternatives where there are fewer interpretable diagnostics available. The remaining results for other scRNA datasets are shown in Appendix.

eSVD appropriately adjusts for the library-size using a dimension-reduction smoothing approach

Many authors have concluded that appropriately adjusting for the library-size of each cell is one of the most influential aspects of an effective DE approach [32, 33]. This is critical for scRNA data since we do not wish to deem genes as DE simply because certain cells have a larger sequencing depth than others. Instead, genes should be deemed as DE if the case and controls have significantly different expression rel-

ative to the cells' sequencing depth. Accounting for this sequencing depth has been the source of numerous debate on how to best normalize the cells' expression [34]. The most commonly-used approach is to log-normalize the gene expressions, but many existing work has cited that this normalization distort qualitative aspects of the scRNA dataset [9, 29].

In order to quantify how well the library-size effect was removed from the scRNA data, we perform diagnostics analogous to those in [29]. Specifically, we group the genes into 6 different bins based on their mean gene expression and observe a significant relation between each gene's expression and the cells' sequencing depth prior to normalizing the scRNA data (Figure 3C). However, after fitting the eSVD, this relation is mostly removed across all bins, discounting the genes with the smallest mean expressions (Figure 3D). Observe that eSVD bears an advantage over other normalization methods such as Scran [35] and SCTransform [29] since eSVD assess the appropriate library-size normalization by accounting for the cells' gene expression and covariate information simultaneously through its dimension-reduction framework, an important benefit when dealing with multi-individual data. Other diagnostics inspired by [29], as well as their comparison to other normalization methods, are included in Appendix.

Lastly, recall that our eSVD framework adjusts the fitted gene expression values by their posterior Negative Binomial distribution. This adjustment relies on first quantifying the dispersion of each gene, i.e., a higher dispersion means the gene's expression patterns conforms less to the learned low-dimensional embedding. We hypothesize a negative correlation between the sequencing-depth normalization and the amount of dispersion, In line with many previous work that investigate this relation [29]. This is because in our setting, a smaller dispersion means the eSVD's fitted values are a good approximation of the relative gene expressions, which often occurs for densely-observed genes. We observe this phenomenon for the lung dataset (Figure 3E), but also other datasets shown in Appendix. Such a scatterplot also enables biologists to broadly survey amount of shrinkage across all the genes. Here, we highlight RPS24 and SMAD3 as examples of genes that highly and lowly shrunk via the posterior distribution since these genes were implicated in previous studies [36, 37, 38]. All these aspects collectively support the claim that eSVD is appropriately adjusting each gene's expression by the sequencing depth, which enables us to investigate the performance of the DE analyses downstream.

eSVD recovers reproducible differences between case and control across multiple datasets

While it is difficult the assess the validity of DE genes in multi-individual scRNA data due to the infeasibility of negative control genes, we hypothesize that a reliable proxy to assess the quality of DE method is to collect two different multi-individual scRNA datasets of the same system and diseases/disorder and see if genes reported to be significant in one dataset display similar significance in the other dataset. Towards this end, we investigate various pairs of datasets but focus on 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) [14]. Since we

are primarily interested in comparing eSVD to other DE methods between this pair of datasets, we do not deploy a multiple-testing procedure to select DE genes, but simply investigate the sets genes with largest test statistic magnitudes of the same cardinality to those reported by the authors for meaningful comparisons. For starters, the volcano plot on the Adams dataset shows that 26 of the 92 genes with largest test statistics derived from our eSVD procedure intersect with the 92 DE genes reported by the authors, resulting in Fisher's exact test p-value of 5.7×10^{-30} (Figure 4B). Similarly, the volcano plot on the Habermann dataset shows that 34 of the 161 genes with largest test statistics derived from our eSVD procedure intersect with the 92 DE genes reported by the authors, resulting in Fisher's exact test p-value of 7.9×10^{-25} (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 broadly speaking [30]. We see this phenomenon demonstrated through the volcano plots (Figure 4B-C).

When comparing among different DE methods, we ask if the leading genes derived from the Adams dataset using eSVD are either (1) genes reported by authors of the Habermann dataset or (2) themselves the leading genes derived from the Habermann dataset using eSVD, 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 is more consistently recovering reproducible signals across both datasets. We focus on pseudo-bulk methods whereby cells of each individual are aggregated into "bulk" expressions prior to performing DESeq2 [7] for this investigation, but include other DE methods for general scRNA data like MAST [28] and SCTransform [29] or specifically multi-individual scRNA data like IDEAS [22] in the in Appendix. We highlight this pseudo-bulk procedure specifically since this procedure has been shown to be more reliable for scRNA with biologically-replicate samples, but its reliability for multi-individual data was not investigated [3]. For our comparisons, we select the 238 genes with the largest test statistics in magnitude from either the Adams or Habermann dataset using either eSVD or DESeq2, since the 92 reported DE genes for the Adams dataset and 161 reported DE genes for the Habermann dataset yield 238 unique genes. The upset plot shows that eSVD indeed yields larger intersections between the two datasets compared to DESeq2 (Figure 4D).

Additionally, we hypothesize that if a DE method recovers reproducible signals, then the test statistics for DE genes should be positively correlated between the two datasets. Certainly, if a reported DE gene has a positive log fold change in one dataset, the other dataset should also show evidence of a positive log fold change. On the other hand, we hypothesize that genes unrelated to the disease/disorder should be uncorrelated test statistics between the two datasets. This 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 relations for eSVD's test statistics (Figure 4E). In contrast, the correlation for DESeq2's test statistic among the reported DE genes is near-zero (Figure 4F). As demonstrated by our previous synthetic simulations, we suspect all the above phenomenons are likely driven by pseudo-bulk methods' lack of accounting for the within-individual variability.

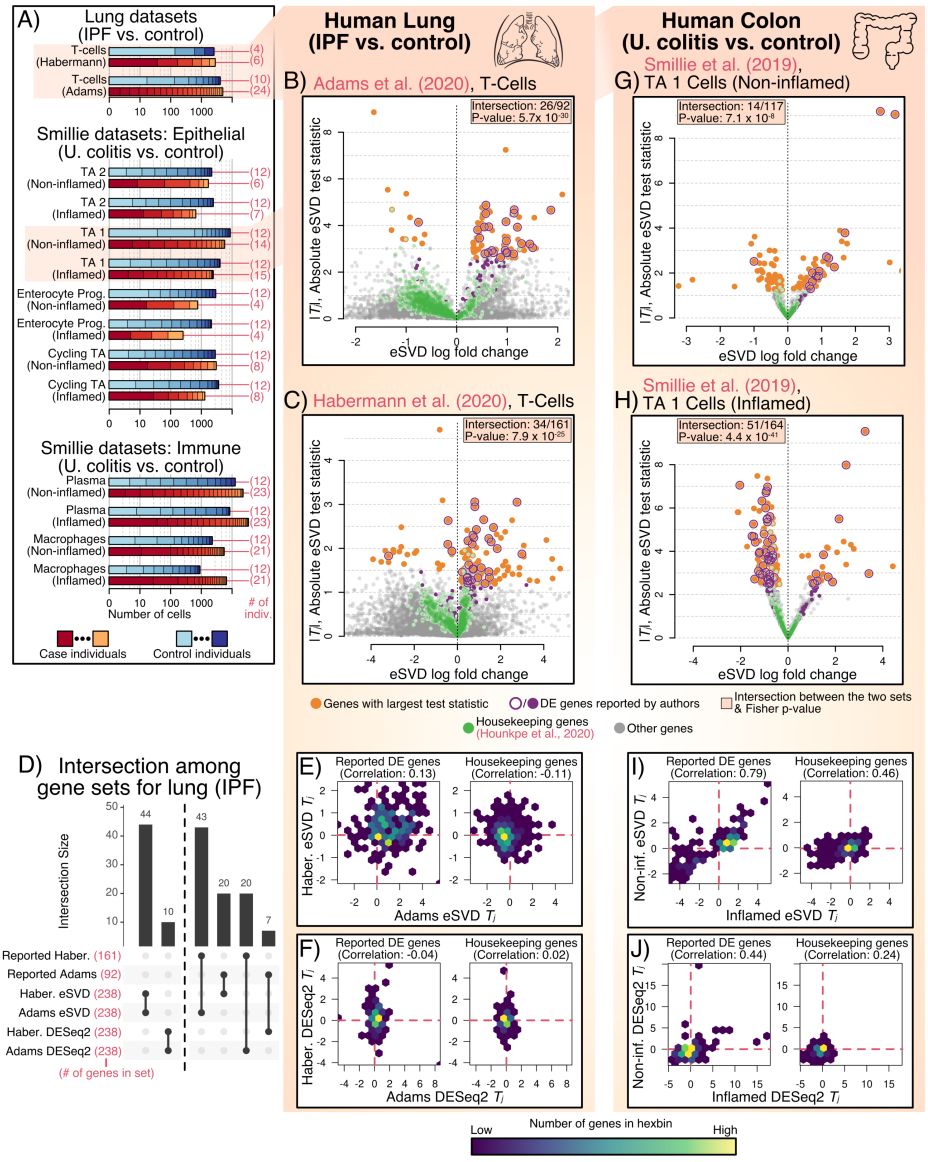

Figure 4 **A** Number of case and control subjects and the total number of cells across cell types and datasets. The x-axis denotes the total number of cells on a log scale, while the partitioning of case and control cells denotes the number of cells from each individual relative to the total number 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'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 or DESeq2. **E,I** Hexplots showing the correlation between genes' eSVD 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.

To further support our empirical claims regarding eSVD, 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 [15]. Each individual with UC contributed cells from inflamed and non-inflamed colon

biopsies, and each healthy individual contributed two biologically-replicate samples from biopsies in analogous colon regions. In our analysis, we treat sample as a different "individual" and perform two DE analysis – 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 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 on 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 [15]. We hypothesize that a higher-powered DE analysis of multi-individual scRNA data should reveal a strong correlation even when focusing on only one cell-type. Indeed, we this positive correlation among eSVD's test statistics ($\rho = 0.79$, Figure 4I). In fact, the test statistic for many genes are larger for in the inflamed analysis than the non-inflamed analysis, suggesting a monotonic development trend among the DE genes as the inflammation worsens. Additionally, while we we see a positive correlation among the housekeeping genes ($\rho = 0.46$), many of such genes have a near-zero test statistic in the non-inflamed analysis. This suggests 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 pseudo-bulk data, the correlation among the reported DE genes is substantially lower (Figure 4J).

eSVD detects DE genes highly enriched with previously-annotated genes

Having demonstrated all the advantages of eSVD, we hypothesize that our method leads 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 [16], henceforth called the "Velmeshev" dataset (Figure 5A). We focus on this particular system since the Simons Foundation Autism Research Initiative (SFARI) maintains routinely curates a list of autism risk genes based on genetic studies from many publications [39], which can provide some semblance of "ground-truth" to compare our eSVD results against. Additionally, a recent currently unpublished bulk-DE analysis of ASD provides an independent list of genes to compare against. We do not expect any method deployed on the Velmeshev dataset to agree with these lists because our analysis is cell-type specific.

When analyzing the 12,984 cells in Layer 2/3, we observe that eSVD estimates 352 DE genes (among when 7094 genes used in the analysis) using a Bayesian FDR cutoff of 0.05, where we used an empirical null multiple-testing procedure to account

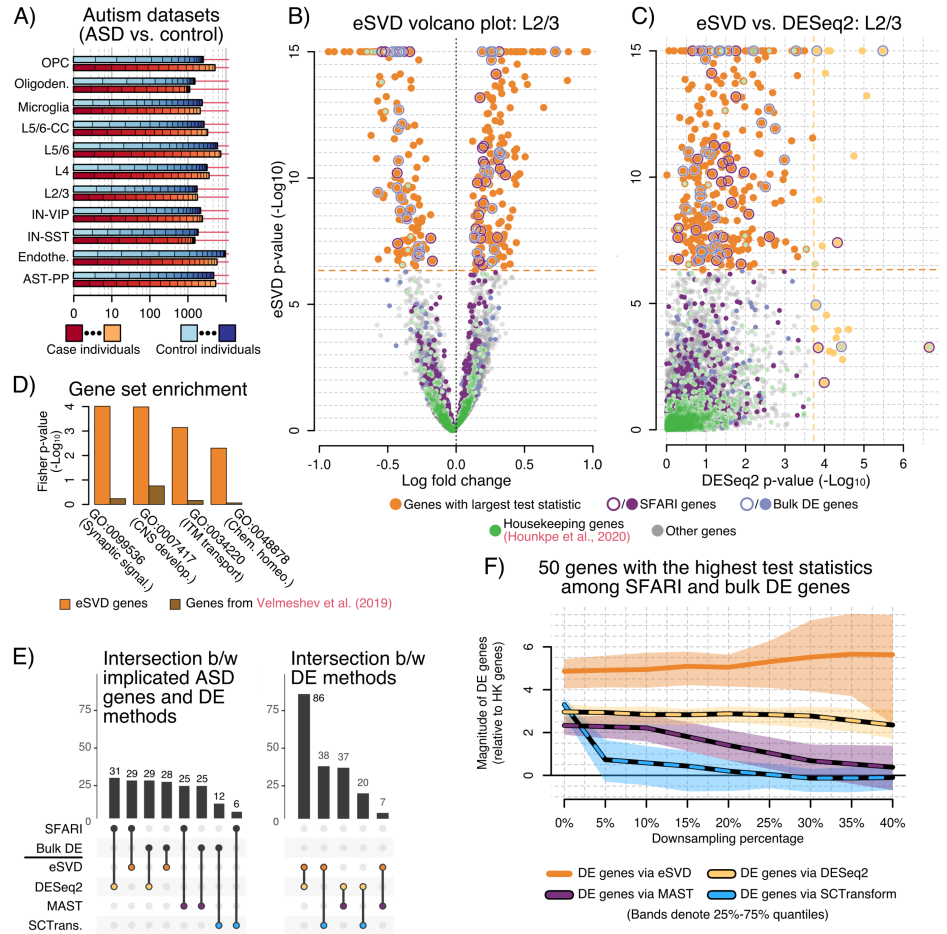

Figure 5 **A** Number of case and control subjects and the total number of cells across cell types for the Velmeshev dataset, analogous to Figure 4**A**. **B** Volcano plot showing eSVD's results for Layer 2/3 cells, where the set of genes exceeding the Bayesian FDR cutoff are depicted in orange. The SFARI genes, bulk DE genes and housekeeping genes are in purple, blue and green respectively. **C** Comparison of the eSVD p-values against the DESeq2 p-values, and the genes exceeding an FDR cutoff for DESeq2 are shown in yellow. **D** Increase in enrichment for relevant GO terms among the 352 DE genes found by eSVD compared to the 109 genes originally reported by the authors. **E** Upset plot comparing the SFARI or bulk DE genes with the DE genes estimated by eSVD, 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 bulk DE genes as the dataset is artificially downsampled.

for potential misspecification [40] (Figure 5B). 34 and 32 of these genes are among the SFARI and bulk-DE genes respectively (among the total 1023 and 1518 genes respectively). While this does not yield a significant Fisher's exact test p-value, we would not have expected this as our analysis is specific to a cell-type. Additionally, when we compare against DESeq2 for the same cells to find DE genes, only 3 and 4 genes overlapped with the SFARI and bulk-DE lists respectively (Figure 5C). This demonstrates that eSVD is able to find more genes compared to DESeq2. We report analogous plots when comparing against other methods such as MAST and SCTransform in the Appendix. 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 352 DE genes found by eSVD are also much more enriched in GO terms that are plausibly related to ASD when compared to the original genes reported for Layer 2/3 [16] (Figure 5D). All these observations suggest that eSVD is able to well-equipped to uncover novel ASD genes.

Next, we investigated for qualitative differences among the methods when compared to the SFARI and bulk-DE genes or each other. In one investigation, we enable easy comparison by finding the top 300 genes with smallest p-values for each method. We then counted how many of these 300 genes intersected with the SFARI or bulk-DE genes (Figure 5E). eSVD and DESeq2 had the highest overlap with both sets (between 28 to 31) when compared to MAST and SCTransform. In fact, eSVD and DESeq2 had an overlap of 82 genes. The observations from Figure 5D and E combined suggests that eSVD inherits many of the advantages of pseudo-bulk methods that other single-cell-level DE methods lack, while improving over pseudo-bulk methods by accounting for within-individual variability. In the second investigation, we asked how stable each method was as the dataset was gradually downsampled (i.e., sampled to become sparser, see Appendix for details). Certainly, 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 most largest Z-scores in magnitude among the SFARI and bulk-DE genes for each method, and ask how their mean Z-scores (relative to the housekeeping gene's Z-scores) 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. We observe that the downsampling percentage increases (i.e., the signal becomes fainter), MAST and SCTransform lose their ability to distinguish between formally highly-differentiable genes and the housekeeping genes. This makes sense, as both 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 are surprisingly stable – they are able to demonstrate a clear separation between the 50 DE genes and the housekeeping genes even when the data is downsampled at 40%. This is also sensible, as DESeq2 aggregates cells among individuals and eSVD pools information between genes, both strategies yielding robustness against sparsity. eSVD is still preferred over DESeq2 here though, as the separation between the DE genes and housekeeping genes is much higher for eSVD.

Conclusions

We demonstrate the nuances of performing cohort-level differential testing for single-cell RNA-seq data. The difficulty primarily stems from the presence of individual-level confounding covariates, which can be difficult to remove using typical DE strategies of regressing out their effects gene-by-gene. Instead, eSVD pools the information across genes in order to remove the confounding covariates, yielding empirical performance that is more promising than current pseudo-bulk DE methods or DE methods currently used for scRNA data when there are no prevalent individual-level covariates present. We achieve this through a matrix factorization strategy, where we estimate both the coefficients associated with the covariates as

well as each cell's and gene's latent vectors. We then shrink the estimated denoised gene expressions via the posterior mean according the Gamma-Poisson distribution. This helps dampen potential over-smoothing effects induced by the matrix factorization. We then deploy a test-statistic designed to test for DE on the cohort-level instead of the cellular-level. We hope that eSVD would not only be beneficial for inspiring cohort-level DE tests for future single-cell assays beyond scRNA, but also single-cell eQTL analyses where there are abundant individual-covariate effects that need to be properly removed.

Author details
[1]Department of Statistics & Data Science, University of Pennsylvania, Philadelphia, 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. 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)
3. 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. bioRxiv (2021)
4. 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)
5. 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)
6. 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)
7. 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)
8. Barry, T., Wang, X., Morris, J.A., Roeder, K., Katsevich, E.: SCEPTRE improves calibration and sensitivity in single-cell CRISPR screen analysis. Genome biology **22**(1), 1–19 (2021)
9. 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)
10. 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)
11. Andrews, T.S., Hemberg, M.: False signals induced by single-cell imputation. F1000Research **7** (2018)
12. 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)
13. 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)
14. 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)
15. 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)
16. 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)
17. 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)
18. 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)
19. Sarkar, A., Stephens, M.: Separating measurement and expression models clarifies confusion in single cell rna-seq analysis. Nature Genetics **53**(6), 770–777 (2021)
20. 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)
21. 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)

22. 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)
23. 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)
24. 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)
25. Liu, Y., Wang, N., Adams, T.S., 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. bioRxiv (2022)
26. Junttila, S., Smolander, J., Elo, L.L.: Benchmarking methods for detecting differential states between conditions from multi-subject single-cell RNA-seq data. bioRxiv (2022)
27. 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)
28. 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)
29. 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)
30. 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)
31. Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I.W., Ng, L.G., Ginhoux, F., Newell, E.W.: Dimensionality reduction for visualizing single-cell data using UMAP. Nature Biotechnology **37**(1), 38 (2019)
32. 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)
33. 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)
34. 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)
35. 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)
36. 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)
37. 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)
38. 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)
39. gene database, S.: Https://gene.sfari.org/. `https://gene.sfari.org/` Accessed 2022-10-20
40. Efron, B.: Microarrays, empirical bayes and the two-groups model. Statistical science **23**(1), 1–22 (2008)
41. 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)
42. Boyeau, P., Regier, J., Gayoso, A., Jordan, M.I., Lopez, R., Yosef, N.: An empirical bayes method for differential expression analysis of single cells with deep generative models. bioRxiv (2022)