Normalization of Single Cell RNA Sequencing Data Using both Control and Target Genes

Mengjie Chen[1, 2] and Xiang Zhou[3, 4]

1. Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599
2. Department of Genetics, University of North Carolina, Chapel Hill, NC 27599
3. Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109
4. Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109

Corresponding author email: mengjie@email.unc.edu, xzhousph@umich.edu

Abstract

Single cell RNA sequencing (scRNAseq) technique is becoming increasingly popular for unbiased and high-resolutional transcriptome analysis of heterogeneous cell populations. Despite its many advantages, scRNAseq, like any other genomic sequencing technique, is susceptible to the influence of confounding effects. Controlling for confounding effects in scRNAseq data is thus a crucial step for proper data normalization and accurate downstream analysis. Several recent methodological studies have demonstrated the use of control genes for controlling for confounding effects in scRNAseq studies; the control genes are used to infer the confounding effects, which are then used to normalize target genes of primary interest. However, these methods can be suboptimal as they ignore the rich information contained in the target genes. Here, we develop an alternative statistical method, which we refer to as scPLS, for more accurate inference of confounding effects. Our method is based on partial least squares and models control and target genes jointly to better infer and control for confounding effects. To accompany our method, we develop a novel expectation maximization algorithm for scalable inference. Our algorithm is an order of magnitude faster than standard ones, making scPLS applicable to hundreds of cells and hundreds of thousands of genes. With extensive simulations and comparisons with other methods, we demonstrate the effectiveness of scPLS. We apply scPLS to analyze three scRNAseq data sets to further illustrate its benefits in removing technical confounding effects as well as for removing cell cycle effects.

Author Summary

Data normalization is crucial for accurate estimation of gene expression levels and successful down-stream analysis in single cell RNA sequencing (scRNAseq) studies. We present a novel statistical method that solves a key challenge in data normalization for scRNAseq: controlling for the hidden confounding factors (e.g. batch effects, cell cycle effects etc.) and removing unwanted variation. Compare to some recent methods using a small set of control genes to infer and control for confounding effects, we propose instead modeling both control and non-control genes jointly. Through extensive simulations and case studies, we demonstrate that joint modeling enables much more accurate data normalization than previous approaches.

Introduction

Single-cell RNA sequencing (scRNAseq) has emerged as a powerful tool in genetics and genomics. It provides unprecedented insights into many basic biological questions by accurately quantifying gene expression levels at both the single cell resolution and the genome-wide scale. Recently, scRNAseq has been applied to classify novel cell subtypes [1, 2] and cellular states [3, 4], reconstruct cell lineage and quantify progressive gene expression during development [5-8], perform spatial mapping and re-localization [9, 10], identify differentially expressed genes and gene expression modulars [11-13], and investigate the genetic basis of gene expression variation by detecting heterogenic allelic specific expressions [14, 15].

Like any other genomic sequencing experiment, scRNAseq studies are influenced by many factors that can introduce unwanted variation in the sequencing data and confound the down-stream analysis [16]. Due to low capture efficiency and low amount of input material, such unwanted variation are exacerbated in scRNAseq experiments [17]. As a result, adjusting for confounding factors and normalizing scRNAseq data is crucial for accurate estimation of gene expression levels and successful down-stream analysis [16-20]. However, depending on the source, adjusting for confounding factors in scRNAseq can be non-trivial. Some confounding effects, such as read sampling noise and drop-out events, are direct consequences of low sequencing-depth, which are random in nature and can be readily addressed by probabilistic modeling using existing statistical methods [18-22]. Other confounding effects are inherent to a particular experimental protocol and can cause amplification bias, but can be easily mitigated by using new protocols [23]. Yet other confounding effects are due to observable batches and can be adjusted for by including batch labels and technician ids as covariates. However, many confounding factors are hidden and are difficult or even impossible to measure. Common hidden confounding factors include various technical artifacts during library preparation and sequencing, and unwanted biological confounders such as cell cycle status. These hidden confounding factors can cause systematic bias, are notoriously difficult control for, and are the focus of the present study.

To mitigate the influence of hidden confounding factors, existing statistical methods in scRNAseq studies take advantage of a set of control genes [24, 25]. Specifically, these methods divide genes into two sets: a control set of genes that are used to infer the confounding factors and a target set of genes that are of primary interest. The confounding factors inferred from the control set are used to remove unwanted variation in the target genes for subsequent downstream analysis. For example, most scRNAseq studies add ERCC spike-in controls during the PCR amplification and sequencing steps. The spike-in controls can be used to capture the hidden confounding technical noise associated with the experimental procedures for normalizing genes of primary interest [26]. Similarly, most scRNAseq studies include a set of control genes that are known to have varying expression levels across cell cycles. These cell cycle genes can be used to capture the unmeasured cell cycle status of each cell, which is further used to normalize target genes [24]. Various statistical methods can be used to infer the confounding effects from control genes. For example, principal component analysis (PCA) or factor models extract the principal components or factors from the set of control genes as surrogates for the confounding factors [25, 27-29]. Linear mixed models (LMMs) or single cell latent variable models (scLVM) construct a sample relatedness matrix based on the control genes to capture the influence of the confounding factors [24, 26, 30]. Other methods fit smooth curves to interpolate and estimate the overall contribution of the technical variation from the control genes [31].

Although straightforward, these statistical methods overlook one important fact -- that the hidden confounding factors not only influence the control genes but also the target genes – i.e. the exact reason that we need to remove such unwanted variation in the first place. Because the confounding factors influence both control and target genes, using control genes alone to infer the confounding factors can be suboptimal as it fails to use the information from target genes. Thus, existing statistical methods do not make full use of the information contained in the data to remove the unwanted variation that can influence downstream analysis.

To infer latent confounding factors from scRNAseq studies and remove unwanted variation, we develop a novel statistical method, which we refer to as scPLS. scPLS is based on the partial least squares regression models and incorporates both control and target genes to infer hidden confounding effects. In addition, our method can model other systematic biological variation and heterogeneity, which are often observed in the target genes. By incorporating such

systematic heterogeneity, we can further improve the estimation of the confounding factors and the removal of unwanted variation. To make our method widely applicable, we also develop a novel efficient estimation algorithm that is scalable to thousands of cells and tens of thousands of genes. With simulations and three real data examples, we illustrate the effectiveness of our method for gene expression normalization and accurate down-stream analysis in scRNAseq studies.

Results

Method Overview

We consider dividing genes into two sets: a control set that contains control genes and a target set that contains genes of primary interest. The control genes are selected based on the purpose of the analysis. For example, the control set would contain ERCC spike-ins if we want to remove technical confounding factors, and would contain cell cycle genes if we want to remove cell cycle effects. We use the partial least squares (PLS) regression to jointly model both control and target genes:

$$\boldsymbol{x}_i = \boldsymbol{\Lambda}_x \boldsymbol{z}_i + \boldsymbol{\epsilon}_{xi}, \boldsymbol{\epsilon}_{xi} \sim N(0, \boldsymbol{\Psi}_{\mathrm{xi}})$$

$$\boldsymbol{y}_i = \boldsymbol{\Lambda}_y \boldsymbol{z}_i + \boldsymbol{\Lambda}_u \boldsymbol{u}_i + \boldsymbol{\epsilon}_{yi}, \boldsymbol{\epsilon}_{yi} \sim N(0, \boldsymbol{\Psi}_{\mathrm{yi}})$$

where for *i*'th individual cell, $\boldsymbol{x}_i$ is a *q*-vector of expression level for *q* control genes; $\boldsymbol{y}_i$ is a *p*-vector of expression level for *p* target genes; $\boldsymbol{z}_i$ is k_z-vector of unknown confounding factors that affect both control and target genes; the confounding effects are represented by the *q* by k_z loading matrix $\boldsymbol{\Lambda}_x$ for the control genes and the *p* by k_z loading matrix $\boldsymbol{\Lambda}_y$ for the target genes; $\boldsymbol{u}_i$ is a k_u-vector of unknown factors in the target genes representing structured variation (see below); $\boldsymbol{\Lambda}_u$ is a *p* by k_u loading matrix; $\boldsymbol{\epsilon}_{xi}$ is an *q*-vector of idiosyncratic error with covariance $\boldsymbol{\Psi}_{\mathrm{xi}} = diag(\sigma_{x1}^2, \cdots, \sigma_{xq}^2)$; $\boldsymbol{\epsilon}_{yi}$ is an p-vector of idiosyncratic error with covariance $\boldsymbol{\Psi}_{\mathrm{yi}} = diag(\sigma_{y1}^2, \cdots, \sigma_{yp}^2)$. We have assumed that the expression levels of each gene have been centered to have mean zero, allowing us to ignore the intercept.

Our model includes two types of unknown latent factors. The first set of factors, $\boldsymbol{z}_i$, represents the unknown confounding factors that affect both control and target genes. The effects of $\boldsymbol{z}_i$ on the control and target genes are captured in the loading matrices $\boldsymbol{\Lambda}_x$ and $\boldsymbol{\Lambda}_y$, respectively. We call $\boldsymbol{z}_i$ the confounding factors throughout the text, and we aim to remove the factor effects $\boldsymbol{\Lambda}_y \boldsymbol{z}_i$ from the target genes. The second set of factors, $\boldsymbol{u}_i$, aims to capture a low dimensional structure of the expression level of *p* target genes. The factors $\boldsymbol{u}_i$ are sometimes referred to as gene signatures, representing intermediate factors that coordinately regulate a set of genes in biological processes and/or specific experimental perturbations. These factors can be interpreted as cell subtypes, treatment status, transcription factors or regulators of biological pathways in different studies [32-36]. Although $\boldsymbol{u}_i$ could be of direct biological interest in many data sets, we do not explicitly examine the inferred $\boldsymbol{u}_i$ here. Rather, we view modeling $\boldsymbol{u}_i$ in the target genes as a way to better capture the complex variance structure, and to facilitate precise estimation of the confounding factors $\boldsymbol{z}_i$. For simplicity, we call $\boldsymbol{u}_i$ the biological factors throughout the text, though we note that $\boldsymbol{u}_i$ could well represent non-biological processes such as treatment or environmental effects. Thus, the expression levels of the control genes can be described by a linear combination of the confounding factors $\boldsymbol{z}_i$ and residual errors; the expression levels of the target genes can be described by a linear combination of the confounding factors $\boldsymbol{z}_i$, the biological factors $\boldsymbol{u}_i$ and residual errors. For both types of confounding factors, we are interested in inferring the factor effects $\boldsymbol{\Lambda}_y\boldsymbol{z}_i$ and $\boldsymbol{\Lambda}_u\boldsymbol{u}_i$ rather than the individual factors $\boldsymbol{z}_i$ and $\boldsymbol{u}_i$. Therefore, unlike in standard factor models, we are not concerned with the identifiability of the factors. Figure 1 shows an illustration of scPLS.

We develop an expectation-maximization (EM) algorithm to estimate the parameters in the latent factor model. Our algorithm treats the latent factors as missing data and uses an iterative procedure to compute the expectation of factors in the E-step and update the factor loading matrices in the M-step (details in the Supplementary Text). This naïve EM algorithm, however, is computationally expensive, and scales quadratically with the number of genes and linearly with the number of samples. To improve the computational speed, we develop a new EM-in-chunks algorithm. Our algorithm is based on the observation that the expression levels of the target genes are determined by the same set of underlying factors and that these factors can be estimated accurately even with a small subset set of target genes. This allows us to randomly divide target genes into dozens of chunks, compute the expectation of the factors in each chunk separately in the E-step, and then average these expectations across chunks. With the averaged expectations, we then update the factor loading matrices in the M-step. Thus, our new algorithm modifies the E-step in the naïve algorithm and is *c* times faster than the naïve one, where *c* is the number of chunks. Simulations show that our EM-in-chunks algorithm yields comparable results to the naïve EM algorithm with respect to estimation errors, but can be an order of magnitude faster (Table S1). To determine the number of confounding factors and biological factors, we evaluate the likelihood on a grid of k_z and k_z values and choose the optimal combination that minimizes the Bayesian information criterion (BIC). After parameter estimation, we use the residuals $\widehat{\boldsymbol{y}_l} = \boldsymbol{y}_i - \widehat{\boldsymbol{\Lambda}}_y\widehat{\boldsymbol{z}_l}$ as the de-noised values for subsequent analysis. Notice that the residuals are only free of confounding effects $\boldsymbol{\Lambda}_y\boldsymbol{z}_i$ but still contain the biological factor effects $\boldsymbol{\Lambda}_u\boldsymbol{u}_i$.

Simulations

Our method improves upon previous methods in two important ways. First, it explicitly models the effects of confounding factors $\boldsymbol{z}_i$ on both control and target genes. Second, it accounts for a low dimensional structure in the expression levels of target genes with the biological factors $\boldsymbol{u}_i$. Both features are expected to improve the normalization of target genes. To illustrate the benefits of these two features, we perform a simulation study. Details of the simulations are described in Methods. Briefly, we simulate gene expression levels for 50 control genes and 1,000 target genes for 200 cells. These 200 cells come from two equal-sized groups, representing two treatment conditions or two sub-cell types. Among the 1,000 target genes, only 100 of them are differentially expressed (DE) between the two groups and thus represent the signature of the two groups. The effect sizes of the DE genes are simulated from a normal distribution, and we scale the effects further so that the group label explains a fixed percentage of phenotypic variation (PVE) in expression levels in the DE genes (ranging from 1% to 20%, with 1% intervals). In addition to the group effects, we simulate the confounding factors $\boldsymbol{z}_i$ to explain 10% PVE in either the control or the target genes, the biological factors $\boldsymbol{u}_i$ to explain either 0% (i.e. no effect) or 30% PVE of the target genes, and the residual errors to explain the rest of PVE. To make the simulations realistic, these PVE values are based on real data sets. For the confounding factors $\boldsymbol{z}_i$, we consider two scenarios: a simple scenario where $\boldsymbol{z}_i$ is not correlated with the group label, and a complicated but perhaps more realistic scenario where $\boldsymbol{z}_i$ is correlated with the group label.

We then compare our method to two commonly used methods -- the PCA method and the LMM method – that use control genes to infer confounding effects. To illustrate the benefit of the biological factors $\boldsymbol{u}_i$, we also compare our full model to a reduced model that does not contain the biological factors. Our goal on the simulated data is twofold: we want to identify these differentially expressed genes and to classify the 200 cells into two groups. Therefore, we compare the performance of various methods based on two criteria: the power to identify the DE genes and the power to classify cells into two groups. We permute group labels to construct an empirical null and compare methods based on either power given a certain false discovery rate (FDR) for identifying DE genes or accuracy for classification.

Because our method infers the confounding factors $\boldsymbol{z}_i$ using information from both control and target genes, it can estimate the confounding effects more precisely than other methods and thus result in better gene expression estimates in the target set (Figure S1). For instance, the correlation between the true and estimated confounding effects is on average 0.84 for scPLS, 0.81 for PCA, and 0.61 for LMM (p-value < 10^{-16} for scPLS vs PCA and p-value = 0.015 for scPLS vs LMM). Due to the better normalization of expression levels, our method is more powerful than both PCA and LMM in identifying DE genes (Figures 2a and S2). Specifically, when PVE=10%, scPLS achieves an average power of 58.4% for detecting DE genes, while LMM and PCA achieve an average power of 38.5% and 42.2%, respectively; when PVE=20%, scPLS achieves an average power of 71.4% for detecting DE genes, while LMM and

PCA achieve an average power of 55.7% and 58.9%, respectively. The improvement of scPLS over other methods is especially significant in the scenario where the confounding factors are correlated with the group label, as the benefits of using the target genes in addition to the control genes become more important (Figure 2a vs Figure S2). Interestingly, for detecting DE genes, when the group effect size is small, methods that remove technical confounding factors may perform even worse than uncorrected data, presumably because of model over-fitting (Figure 2a). However, while PCA and LMM outperform the uncorrected data only when the group effect size is above PVE=10%, scPLS outperforms the uncorrected data for a small PVE of 5%, suggesting that scPLS is more robust to model over-fitting.

One important benefit of scPLS, compared with others, is that it is less sensitive to the number of control genes used in the analysis. Because scPLS does not completely rely on information from the control genes, it achieves reasonably good performance even if we only use a much smaller subset of control genes. In contrast, the performance of other methods compromise more quickly with a reduced number of control genes (Figure 2b). For instance, for detecting DE genes, when PVE=20%, using 10 control genes instead of 50 results in an average power reduction of only 3.2% for scPLS, but results in an average power reduction of 7.6% and 7.8% for LMM and PCA, respectively (p-value = $8.21\text{x}10^{-05}$ for scPLS vs PCA and p-value = $4.79\text{x}10^{-05}$ for scPLS vs LMM).

The higher power of scPLS to detect DE genes also translates to a better performance of classifying single cells. To visualize classification performance, we extract the principal components (PCs) from normalized expression levels of the top 100 DE genes and visualize the data on PC1 and PC2. For scPLS, PC1 and PC2 clearly separate cells into the two known cell groups. For both LMM and PCA, the separation is less clear (Figure 2c). To further quantify the classification performance, we apply the support vector machine (SVM) to classify the cells. We perform a five-fold cross-validation, training SVM with 80% of the samples and evaluating the prediction accuracy with the rest of the samples. As expected, scPLS outperforms the other two methods (Figure 2d). When PVE is 10%, scPLS achieves an average accuracy of 98% across 10 replicates, while LMM and PCA only achieve 84.3%. When PVE is 20%, scPLS achieves 99%, while LMM and PCA achieve 91.5% and 88.5%, respectively.

Finally, our full model also compares favorably with the reduced model where we do not model biological factors. In particular, the full model performs better than the reduced model in the presence of biological factors, but does not perform much worse when there are no biological factors in the data (Figure S3a). For instance, when biological factors are included in the simulated data, the reduced model achieves 34.8%, compared with 58.4% from the full model when PVE=10%. When PVE=20%, the reduced model only achieves a power of 52.1% for detecting DE gene, compared with 71.4% from the full model. As a consequence, the reduced model achieves an average prediction accuracy of 94.8% in classification, compared with 98% from the full model when PVE=10%. On the other hand, when there are no biological factors in then simulated data and when PVE=10%, the reduced model achieves a power of 51.7%, while the full model still achieves a power of 50.4% (using 2 biological factors) -- a very small reduction (Figure S3b). Similarly, when PVE=20% the reduced model achieves a power of 66.2%, while the full model still achieves a power of 65.2%. Importantly, the comparison results are not sensitive with respect to the number of biological factors used in fitting the model when no biological factors are present (Figure S3b). As it is often unknown whether a low-rank structural variation exists in a real data set, our simulation suggests that we can always include the biological factors $\boldsymbol{u}_i$ in the model.

Real Data Applications

Next, we illustrate the benefits of our method in three real data sets. The first dataset is used to demonstrate the effectiveness of scPLS in removing the technical confounding effects by using ERCC spike-ins. Removing technical confounding effects is a common and important task in transcriptome analysis. The second dataset is used to demonstrate the effectiveness of scPLS in removing cell cycle effects by using a known set of cell cycle genes. Removing cell cycle effects can reveal gene expression heterogeneity that is otherwise obscured. Finally, with the third dataset, we demonstrate the effectiveness of scPLS in removing both technical confounding effects and cell cycle effects.

Removing Technical Confounding Factors

The first dataset consists of 251 samples of mouse embryonic stem cells (mESCs). The cells were cultured in two different media: two-inhibitor (2i) medium or serum medium. The cells in each medium were collected in two different ways: either as single cell samples subject to scRNAseq or as pooled samples subject to pooled RNAseq. Therefore, the samples can be divided into four different categories based on the medium type and collection type. We use ERCC spike-ins as controls and all other genes as the target. In this data, scPLS infers $k_z = 1$ confounding factors and $k_u = 1$ biological factors. In the target genes, the confounding factors and biological factors explain a median of 16.1% and 33.9% of gene expression variance, respectively. The PVE by the technical factors are higher in highly variable genes. For example, in the top 500 genes with the largest variance, the two sets of factors explain a median of 18.6%

and 32.4% of gene expression variance, respectively. The PVE by the confounding and biological factors can be as high as 70.0% and 87.1% in the target genes (Figure 3a).

We first visualize the uncorrected data and data corrected by scPLS with heatmap (Figure 3b). Before correction, there is a large variation in gene expression levels across samples in each category. However, this large variance is due to technical confounding. After correction, the variability of samples within each category is considerably reduced.

We next compare the efficacy of scPLS with other methods in removing technical confounding factors. Here, in addition to LMM and PCA, we also compare with RUV, a method specifically designed to remove technical confounding factors by using control genes [25]. Because the data consists of four sample categories, we compare the performance of methods based on whether the corrected data can better reveal four distinct groups. In particular, if a method effectively removes confounding effects, the top PCs from the corrected data can be used to better distinguish the four categories. Thus, we extract the top PCs from both corrected data and uncorrected data. For each PC, we compute a ratio of within-category variance and the total-variance. A small ratio suggests that this particular PC can be used to distinguish the four categories. We then contrast the ratio from the k'th PC in the corrected data with the ratio from the k'th PC in the uncorrected data, and compute the reduction in ratio as a measurement of performance; a large reduction suggests better performance. As expected, data corrected by scPLS achieves the largest reduction in ratio, suggesting that scPLS outperforms the other methods (Figure 3c). For example, the within-cluster variance ratio for the first PC is reduced 30% by scPLS. In contrast, the decrease in the ratio for the first PC by PCA, LMM and RUV is only 25.1%, 24.5% and 23.6%, respectively. The trend is apparent for all top PCs. For example, for the 10'th PCs, the ratio is reduced 13.4% by scPLS. In contrast, the ratio reduced by PCA, LMM and RUV are only 10.4%, 9.9% and 9.3%, respectively. The results suggest that scPLS performs better than the alternatives and can better reveal distinct sample categories in the data.

Removing Cell Cycle Effects

The second dataset measured the expression level of 301 single cells from either the utricular or the cochlear sensory epithelia in the inner ear of newborn mice [37]. The cells from each of the two tissues (i.e. utricle and cochlea) can be classified into three cell types (the hair cells, or HCs; the transitional epithelial cells, or TECs; and supporting cells, or SCs), resulting in a total of six distinct cell types. We 282 known cell cycle genes as controls and all other genes as targets. scPLS infers $k_z = 2$ confounding factors of cell cycle and $k_u = 2$ biological factors. The cell cycle confounding factors and biological factors explain a median of 11.9% and 1.4% of gene expression variance, respectively. The PVE by the latent factors are higher in highly variable genes. For example, both factors explain a median of 18.0% and 2.4% of gene expression variance, respectively, in the top 500 genes with the largest variance. The PVE by the confounding and biological factors can be as high as 91.5% and 73.4% (Figure 4a).

We first perform a PC analysis on the uncorrected data and the scPLS corrected data. In the uncorrected data, PC1 and PC2 separated cells into two clusters (Figure 4b). However, these two clusters do not represent tissue groups. Instead, one cluster consists largely of HCs from the two tissues while the other cluster consists of the other four cell types. PC1 and PC3 separate cells into the expected two tissue groups. On the other hand, PC1 and PC2 of scPLS corrected data directly cluster cells according to the cell type (Figure 4b). This suggests that PC1 of the uncorrected data may represent cell cycle effects. To directly test this, we extract PCs from the cell cycle genes and denote the first PC as the cell cycle PC (ccPC). We expect this ccPC to represent the majority of the cell cycle effects, as it accounts for 50.1% of variance among the top five PCs of cell cycle genes. In either the uncorrected data or scPLS corrected data, we calculate the correlation of PC1, PC2 and PC3 with ccPC (Figure 4c). In the uncorrected data, PC1 is highly corrected with ccPC with a striking correlation of 0.96 (p-value < 10^{-16}). PC2 is also correlated with ccPC but to a much lesser degree, and PC3 is not correlated (r= 0.23 and 0.02, p-values = 5.6 x 10^{-5} and 0.74). After scPLS adjusting for cell cycle effects, the correlation with PC1 is almost completely gone (r=6.8 x 10^{-3}, p-value = 0.90). The correlations of ccPC with PC2 are also reduced to near zero (r= 8.4 x 10^{-5}, and 0.04, p-values = 0.99 and 0.55). The correlation results suggest that scPLS can successfully remove cell cycle effects.

We then compare the performance of scPLS with other methods in this data. In addition to PCA and LMM, this time, we also include scLVM for comparison. scLVM is specifically designed to remove cell cycle effects [24]. Because we do not know the true cell cycle label here (unlike the following data; see below), we use two different criteria to compare different methods.

Our first criterion is the correlation of PCs with ccPC as described in the paragraph above. All methods are able to reduce the correlation between PC1 and ccPC. For example, ccPC is not highly correlated with PC1 from PCA (r = 0.02), LMM (r= 0.02) or scLVM (r=0.05). However, among all the methods, PC1 from scPLS corrected data has the smallest correlation with ccPC (r=5.6 x $10^{-5;}$ p-value < 1.0 x 10^{-16} for scPLS vs PCA, p-value < 1.0 x 10^{-16} for scPLS vs LMM, and p-value = 0.001 for scPLS vs scLVM), suggesting that scPLS works best among all. A similar trend is observed for PC2. However for PC2, both scPLS and scLVM work well and reduce the correlation to near zero.

Our second criterion to compare the methods is based on a subset analysis. For this criterion, we randomly split the cell cycle genes into two sets: one set (150 genes) is used as the control set as usual, but the other set (132 genes) is included in the target gene set and used for validation purpose. We perform four replicates of split. We reason that, if a method works well to remove the cell cycle effects, then the 132 genes in the validation set should be free of cell cycle effects, and thus would no longer be highly correlated with each other or with genes in the control set. In the uncorrected data, the control genes and validating genes are highly positively correlated with each other (Figure 4d). The standard hierarchical clustering algorithm cannot separate these two sets and genes from the two sets are intermingled in the hierarchical cluster. In contrast, after scPLC correction, the high correlations between the control genes and validating genes are largely gone. The standard hieratical clustering algorithm can easily separate the control genes and the validating genes (Figure 4d). To quantitatively measure the reduction in correlations, we compute all pair-wise correlations between genes from the controls set and genes from the validation set (Figure 4e). scPLS reduces the median correlation among genes from 0.09 to 0.043 (p-value = 8.0×10^{-5}), and reduces the 90% quantile correlation from 0.190 to 0.108 (p-value = 0.0003). scLVM also reduces correlations, but to a lesser amount compared with scPLS. For example, scLVM reduces the 90% quantile correlation among genes to 0.115, which is significantly less than scPLS (p-value = 0.047). scLVM reduces the median correlation to 0.046, which is also significantly less than scPLS (p-value = 0.017). The results from PCA and LMM are not significantly different compared to scPLS, suggesting that these methods are also effective in removing cell cycle effects.

In conclusion, based both criteria, scPLS can effectively remove cell cycle effects, and more so than the other methods.

Removing Both Technical Effects and Cell Cycle Effects

Our method can also be used to remove both technical effects and cell cycle effects. We can accomplish this in two steps. First, we use the spike-in genes as the control set and all genes as the target set to jointly infer and remove the technical confounding effects. Then, with the corrected gene expression levels from the first step, we further use the cell cycle genes as the control set and other genes as the target set to estimate and remove the cell cycle effects.

To demonstrate the effectiveness of the two-step approach, we apply our method to a third dataset. This dataset contains the transcriptional profile of 182 embryonic stem cells (ESCs) with pre-determined cell-cycle phases (G1, S and G2M) [24]. We used 92 ERCC spike-ins as the control genes and a set of 9,121 genes selected in the previous study [24]. To remove cell cycle effects, we use 629 annotated cell-cycle genes as controls and other genes as targets. scPLS infers $k_{z1} = 1$ technical confounding factors, $k_{z2} = 1$ cell cycle confounding factors, and $k_u = 1$ biological factors at the second step. These factors explain a median of 0.4%, 3.1%, and 0.7% of gene expression variance, respectively (Figure 5a). The PVE by the three components -- the confounding, cell cycle factors and biological factors -- can be as high as 40.0%, 74.8% and 37%. We visualize the uncorrected data and scPLS corrected data with a PCA plot (Figure 5b). In the uncorrected data, there is a clear separation of cells according to cell-cycle stage. Such separation of cells is not observed in the corrected data, indicating that the cell cycle related expression signature is effectively removed.

We compare the performance of scPLS with the other three methods. Here, both LMM and PCA use the same two-step procedure while scLVM uses the procedure developed earlier [24]. We evaluate the performance based on two different criteria. For the first criterion, we compute for each gene the proportion of expression variance explained by the cell cycle factor. We denote this quantity as PVEi, which stands for inferred PVE. Because the cell-cycle stage of each cell has been experimentally determined in this data set, we further compute the variance explained by the true cell cycle labels. We denote this quantity as PVEt, which stands for true PVE. For scPLS, PVEi and PVEt are highly correlated ($r^2 = 0.77$), demonstrating the efficacy of scPLS (Figure 5c). Importantly, the correlation is similar whether we use the naïve EM algorithm, or the EM-in-chunks algorithm with either the full control set or with a subset of 300 controls (Figure S4). The correlation between PVEi and PVEt in scPLS is higher than that from scLVM (r^2=0.68; p-value < 10^{-16}), LMM (r^2=0.53; p-value < 10^{-16}), and PCA (r^2=0.70; p-value < 10^{-16}) (Figure 5c), suggesting that scPLS works better than the other three methods. To examine whether those factors are indeed good predictors for cell cycle, we further build a regression model between the inferred factors and the know cell cycle phases. For each cell cycle phase, we create a binary indicator variable. We then regress these indicator variables on the estimated technical factors to compute an R^2. A higher R^2 suggests better performance. scPLS achieves R^2 coefficients of 0.33, 0.46 and 0.43 for G1, S and G2M, respectively. In comparison, PCA achieves R^2 coefficients of 0.32, 0.42 and 0.40, and LMM achieved 0.28, 0.40, and 0.38.

To further validate the efficacy of our approach, we again randomly split the cell cycle genes into two sets: one is used as the controls (300 genes), and the other (329) as a validation set, included in the target set. In the uncorrected data, the control genes and validating genes are highly correlated with each other and intermingled (Figure S5b). After scPLC correction, the correlation between the control genes and the validating genes is largely gone and the two set could be easily separated. To quantitatively measure the reduction in correlations, we compute all pair-wise correlations between genes from the controls set and genes from the validation set (Figure S5a). scPLS reduces the median correlation among genes from 0.075 to 0.052 (p-value = 0.0001), and reduces the 95% quantile correlation from 0.219 to 0.158 (p-value = 0.0001). As a comparison, PCA and LMM reduce the median correlation among genes to 0.057, and 0.056, both of which are significantly less than scPLS (p-value = 7.0 x 10^{-5}, and 5.53 x 10^{-5} for scPLS vs PCA, scPLS vs LMM, respectively). scLVM reduces the median correlation among genes to 0.052, which is not significantly different from scPLS (p-value =0.483).

In conclusion, scPLS performs better than the other methods based on multiple criteria.

Discussion

We have presented a new method, scPLS, for removing hidden confounding factors and normalizing expression data in scRNAseq studies. scPLS uses both control and target genes to infer the confounding factors and is more effective in removing confounding effects than previous approaches that use control genes alone. With both simulations and applications to three real data sets, we have demonstrated the effectiveness of our method for controlling for hidden confounding effects in scRNAseq studies.

Although we have focused on the applications of our method to scRNAseq studies, our method can be readily applied to other genomic sequencing studies. For instance, our method can be used to remove confounding effects from gene expression levels in bulk RNAseq studies [38] or from methylation levels in bisulfite sequencing studies [39]. The main requirement of our method is a set of pre-specified control genes that are measured together with the target genes in the sequencing studies. It is often straightforward to obtain such control genes. For example, many scRNAseq studies include a set of ERCC spike-in controls that could be used to model and remove technical confounding effects [26]. Even when such ERCC spike-in controls are not present, we can select a known set of house-keeping genes as controls to remove technical confounding [25]. Similarly, we can use a set of known cell cycle genes to remove cell cycle effects. Importantly, unlike other methods, the performance of scPLS is not sensitive to the number of genes included in the control set, and yields comparable results even when a much smaller number of control genes is used (Figure 2b). This is because scPLS not only uses information from control genes but also relies on information from target genes. Insensitivity to the control set makes scPLS especially suited to removing confounding factors in studies where a control set is not clearly defined. Because of its effectiveness and robustness, we expect scPLS to be widely useful in normalizing data from sequencing studies.

One important feature of scPLS is that it includes a low-rank component to model the structured biological variation often observed in real data. By decomposing the (residual) gene expression variation into a low-rank structured component that is likely to be contributed by a sparse set of biological factors, and an unstructured component that reflects the remaining variation, scPLS can better model the residual error structure for accurate inference of confounding effects. Although here we have focused on using the biological factors to better infer the confounding effects, we note that the low-rank biology factors themselves could be of direct interest. In fact, low-rank factors inferred from many data sets using standard factor models have been linked to important biological pathways or transcription factors [32-36]. Inferring the biological factors using our current model is not feasible at the moment, however:

because of model identifiability, scPLS can only be used to infer the biological effects (i.e. $\boldsymbol{\Lambda}_u \boldsymbol{u}_i$) but not the biological factors (i.e. $\boldsymbol{u}_i$). That said, additional assumptions can be made on the structure of the factors or the factor loading matrices to make factor inference possible [40]. For example, we could impose sparsity assumptions on the low-rank factors to facilitate the inference of a parsimonious set of biological factors. Exploring the use of biological factors from scPLS is an interesting avenue for future research.

Like many other methods for scRNAseq [31] or bulk [41, 42] RNAseq studies, scPLS requires a data standardization step that converts the count data into quantitative expression data. Different conversion methods can affect the interpretation of the data and are advantageous in different situations [16]. Because scPLS does not rely on a particular standardization procedure, scPLS can be paired with any conversion methods to take advantage of their benefits. One important feature of scPLS is that it directly models quantitative expression data instead of raw count data. Despite the count nature of sequencing data, it has been show that there is often a limited advantage of modeling the raw read counts directly, at least for RNAseq studies [43, 44]. Statistical methods that convert and model the quantitative expression data have been shown to be robust [41, 42] and most large scale bulk RNAseq studies in recent years have used normalized data instead of count data [28, 45-47]. However, we note that, unlike bulk RNAseq studies, single cell RNAseq data often come with low read depth. In low read depth cases, modeling count data while accounting for over-dispersions (e.g. dropout events) in single cell RNAseq studies may have added benefits [17, 19]. Therefore, extending our framework to modeling count data [48, 49] is another promising avenue for future research.

Methods

scRNAseq Data

We applied our method to three real data sets.

The first dataset contained both single cell sequencing data and bulk sequencing data, with a total of 251 samples cultured in two different media [18]. In particular, it included 74 mouse embryonic stem cells (mESCs) single cells cultured in a two-inhibitor (2i) medium, 45 mESCs single cells cultured in a serum medium, 56 samples with pooled RNA from mESCs cultured in a 2i medium, and 76 samples with pooled RNA from mESCs cultured in serum. The raw UMI counts (kindly provided by the authors) data contained measurements for 92 ERCC spike-ins and 23,459 genes. Due to the low coverage of this dataset (median coverage equals 1), we filtered out lowly expressed genes and selected only genes that had at least five counts in more than 1/3 of the cells. This filtering step resulted in a total of 17 ERCC spike-ins used as the controls and 2,795 genes used as the targets. Following previous approaches [4, 10, 11, 24], we used log10 transformed values for analysis.

The second dataset contained single cell expression data from the utricular and the cochlear sensory epithelia in the inner ear of newborn mice [37]. We obtained the data from GEO website with accession number GSE71982. The original RSEM data contained count measurements for 26,585 genes of 321 cells. We removed cell types that were represented by less than five cells and focused on the resulting 307 cells from a total of six distinct cell types. The six cell types included the hair cells (HCs), the transitional epithelial cells (TECs), and supporting cells (SCs) from each of the two tissues (utricle and cochlea). For genes, we further filtered out lowly expressed genes that had less than five counts in more than 90% of the cells and fewer variable genes that had variance less than 0.75. This yielded a set of 4,128 genes, 282 of which were annotated as cell cycle genes (see below). Among these genes, we used 3,846 non cell cycle as targets and 282 cell cycle genes as controls. Again, we used log10 transformed values for analysis. In the subset analysis, we randomly sampled 150 genes from the cell cycle genes as controls and used the rest 132 genes in the target set. We only performed four replicates in the subset analysis because results were very consistent across replicates.

The third dataset contained the transcriptional profile of 182 embryonic stem cells (ESCs) with 92 ERCC spike-ins [24]. We obtained the data from ArrayExpress database with accession number E-MTAB-2805. The cells were in three different cell-cycle phases (G1, S and G2M) which had been determined experimentally. There were 59 cells in G1 phase, 58 cells in S phase and 65 cells in G2M phase. We extracted the count data for a list of 9,571 genes as before [24]. Among them, we focused our analysis on 629 annotated cell-cycle genes and 8,942 non cell-cycle genes with the largest variance. Again, we used log10 transformed values for analysis. For scPLS, at the first step, we used 92 spike-ins as controls and 9,571 genes as targets to remove the technical effects. The corrected data were obtained by subtracting the part contributed by the latent technical factors from the measured expression. To further remove cell cycle effects, we used 629 cell-cycle genes as controls and used 8,942 non cell cycle genes as the target set in the second step. In the subset analysis, we randomly sampled 329 genes from the controls and added the rest 300 genes to the target set. Four replicates were generated for the subset analysis.

Both the second and third dataset required a pre-specified set of cell cycle genes as controls. We used a set of 892 cell cycle genes following a previous study [24]. This set of cell cycle genes were generated by intersecting two different data sources: all cell cycle genes with GO ID 0007049 [50], and a list of 600 top-ranked genes from CycleBase [51].

Simulations

We simulated gene expression levels for 50 control genes and 1000 target genes for 200 cells. These 200 cells were simulated from two equal-sized groups, representing two treatment conditions or two sub-cell types. Among the 1000 target genes, only 100 of them were differentially expressed between the two groups and thus represented the signature of the two groups. We generated the expression profile according to the following model:

$$\boldsymbol{x}_i = \boldsymbol{\Lambda}_x \boldsymbol{z}_i + \boldsymbol{\epsilon}_{xi},$$

$$\boldsymbol{y}_i = \boldsymbol{\Lambda}_y \boldsymbol{z}_i + (\mathbf{I} - \boldsymbol{D})\boldsymbol{\Lambda}_u \boldsymbol{u}_i + g_i \boldsymbol{\beta} + \boldsymbol{\epsilon}_{yi},$$

where $\boldsymbol{z}_i$ is the k_z-vector of confounding factors shared between $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$; $\boldsymbol{u}_i$ is the k_u-vector of biological factors specific to $\boldsymbol{y}_i$; g_i is a group indicator function that equals 1 if i'th gene belongs to group one and equals 0 otherwise; $\boldsymbol{\beta}$ is a p-vector of DE effects sizes that equals 0 for non-DE genes; $\boldsymbol{D}$ is p by p diagonal indicator matrix of DE genes: jj'th element of $\boldsymbol{D}$ equals 1 if j'th gene is a DE genes while equals 0 otherwise. This way, the expression levels of DE genes can be described by a linear combination of the confounding factors $\boldsymbol{z}_i$, the DE effects and the residual errors. The expression levels of non-DE genes can be described by a linear combination of the confounding factors $\boldsymbol{z}_i$, the biological factors $\boldsymbol{u}_i$ and residual errors. We considered three different simulation scenarios.

In scenario I, the confounding factors were independent of group effects. Here, we set $k_z = 6$ and $k_u = 10$ as motivated by real data analysis. We simulated each element of $\boldsymbol{z}_i$ and $\boldsymbol{u}_i$ from a standard normal distribution. We simulated each element of $\boldsymbol{\Lambda}_x$ from $N(-0.2, \sigma_l^2)$ and each element of $\mathbf{\Lambda}_y$ from $N(0.2, \sigma_l^2)$. Note that $\mathbf{\Lambda}_x$ and $\mathbf{\Lambda}_x$ were simulated differently to capture the fact that the effect sizes of the confounding factors can be different for control and target genes. We simulated each element of $\mathbf{\Lambda}_u$ from $N(0, \sigma_b^2)$. We simulated each element of $\boldsymbol{\epsilon}_{xi}$ and $\boldsymbol{\epsilon}_{yi}$ from a normal distribution $N(0, \sigma_l^2)$. We simulated the 100 DE effects $\beta_j \sim N(0, \sigma^2)$. We set $\sigma_b^2 = 1.2$, $\sigma_l^2 = 0.4$ and $\sigma^2 = 2.4$ to ensure that, in non-DE genes, the confounding factors $\boldsymbol{z}_i$ explain 10% PVE in either the control or the target genes, the biological factors $\boldsymbol{u}_i$ to explain 30% PVE of the target genes, and the residual errors to explain the rest of PVE. We also set σ^2 such that, in DE genes, the group effects explain a fixed proportion of phenotypic variance that ranging from 1% to 20% with 1% intervals. To make the simulations realistic, these PVE values were all based on real data analysis.

In scenario II, the confounding factors were correlated with group effects. The simulations here were largely similar to scenario I, with the only exception of $\boldsymbol{z}_i$. Here, each element of $\boldsymbol{z}_i$ was simulated depending on the group label of i'th sample; the element of $\boldsymbol{z}_i$ was simulated from $N(0, 1)$ when i'th sample belongs to group one, but was simulated from $N(0.5, 1)$ when the sample belongs to group two. This way, $\boldsymbol{z}_i$ is correlated with group label g_i.

Finally, we also considered a scenario III where there were no group effects. The simulations here were largely similar to scenario I, with the only exception that all genes were non-DE, or $\beta_j = 0$.

To evaluate the performance of different methods, we applied t-test to normalized data to identify DE genes. For each simulated data, we also permuted the group label 10,000 times to construct an empirical null distribution of *p*-values. With the empirical null, we computed power of each method for identifying DE genes based on a fixed empirical false discovery rate (FDR).

Other Methods

We compared scPLS with four other competing methods in both simulations and real data applications. The first method is PCA. We used the top *k* PCs from control genes as covariates to remove confounding effects from target genes. Unlike scPLS, PCA cannot infer *k*. Thus, we used *k*=*k*$_z$ as inferred from scPLS. The second method was LMM, implemented in the GEMMA software [52, 53]. For LMM, we used all control genes to construct a relatedness matrix. Then, we fitted a LMM for each gene in turn from the target set to remove the confounding random effects. The third method was RUV [25]. RUV is specifically designed to remove technical confounding factors. Thus, we only applied RUV to the first real data set. The fourth method was scLVM [24]. scLVM is specifically designed to remove cell cycle effects. Thus we applied scLVM to the second and third real data sets.

Software Availability

The scPLS software is a part of Citrus project and is freely available at: http://chenmengjie.github.io/Citrus/.

Acknowledgements

This work was supported by a start-up fund from the University of North Carolina at Chapel Hill to MC and a start-up fund from the University of Michigan to XZ. MC was also supported in part by the National Institutes of Health (NIH) grants R01 GM105785, R01 CA082659 and P01 CA142538. We thank Dr. Dominic Grun for providing the raw read counts of the first dataset.

References

1. Usoskin D, Furlan A, Islam S, Abdo H, Lonnerberg P, Lou D, Hjerling-Leffler J, Haeggstrom J, Kharchenko O, Kharchenko PV, et al: **Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing.** *Nat Neurosci* 2015, **18:**145-153.
2. Zeisel A, Munoz-Manchado AB, Codeluppi S, Lonnerberg P, La Manno G, Jureus A, Marques S, Munguba H, He L, Betsholtz C, et al: **Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq.** *Science* 2015, **347:**1138-1142.
3. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, Mildner A, Cohen N, Jung S, Tanay A, Amit I: **Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types.** *Science* 2014, **343:**776-779.
4. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al: **Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets.** *Cell* 2015, **161:**1202-1214.
5. Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, Desai TJ, Krasnow MA, Quake SR: **Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq.** *Nature* 2014, **509:**371-375.
6. Tang F, Barbacioru C, Bao S, Lee C, Nordman E, Wang X, Lao K, Surani MA: **Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis.** *Cell Stem Cell* 2010, **6:**468-478.
7. Durruthy-Durruthy R, Gottlieb A, Hartman BH, Waldhaus J, Laske RD, Altman R, Heller S: **Reconstruction of the mouse otocyst and early neuroblast lineage at single-cell resolution.** *Cell* 2014, **157:**964-978.
8. Xue Z, Huang K, Cai C, Cai L, Jiang CY, Feng Y, Liu Z, Zeng Q, Cheng L, Sun YE, et al: **Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing.** *Nature* 2013, **500:**593-597.
9. Achim K, Pettit JB, Saraiva LR, Gavriouchkina D, Larsson T, Arendt D, Marioni JC: **High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin.** *Nat Biotechnol* 2015, **33:**503-509.
10. Satija R, Farrell JA, Gennert D, Schier AF, Regev A: **Spatial reconstruction of single-cell gene expression data.** *Nat Biotechnol* 2015, **33:**495-502.
11. Shalek AK, Satija R, Shuga J, Trombetta JJ, Gennert D, Lu DN, Chen PL, Gertner RS, Gaublomme JT, Yosef N, et al: **Single-cell RNA-seq reveals dynamic paracrine control of cellular variation.** *Nature* 2014, **510:**363-+.
12. Kim KT, Lee HW, Lee HO, Kim SC, Seo YJ, Chung W, Eum HH, Nam DH, Kim J, Joo KM, Park WY: **Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells.** *Genome Biol* 2015, **16:**127.
13. Lee MC, Lopez-Diaz FJ, Khan SY, Tariq MA, Dayn Y, Vaske CJ, Radenbaugh AJ, Kim HJ, Emerson BM, Pourmand N: **Single-cell analyses of transcriptional heterogeneity during drug tolerance transition in cancer cells by RNA sequencing.** *Proc Natl Acad Sci U S A* 2014, **111:**E4726-4735.
14. Borel C, Ferreira PG, Santoni F, Delaneau O, Fort A, Popadin KY, Garieri M, Falconnet E, Ribaux P, Guipponi M, et al: **Biased allelic expression in human primary fibroblast single cells.** *Am J Hum Genet* 2015, **96:**70-80.
15. Deng Q, Ramskold D, Reinius B, Sandberg R: **Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells.** *Science* 2014, **343:**193-196.
16. Stegle O, Teichmann SA, Marioni JC: **Computational and analytical challenges in single-cell transcriptomics.** *Nat Rev Genet* 2015, **16:**133-145.

17. Vallejos CA, Marioni JC, Richardson S: **BASiCS: Bayesian Analysis of Single-Cell Sequencing Data.** *PLoS Comput Biol* 2015, **11:**e1004333.
18. Grun D, Kester L, van Oudenaarden A: **Validation of noise models for single-cell transcriptomics.** *Nat Methods* 2014, **11:**637-640.
19. Kharchenko PV, Silberstein L, Scadden DT: **Bayesian approach to single-cell differential expression analysis.** *Nature Methods* 2014, **11:**740-U184.
20. Kim JK, Kolodziejczyk AA, Illicic T, Teichmann SA, Marioni JC: **Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression.** *Nat Commun* 2015, **6:**8687.
21. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW, McElrath MJ, Prlic M, et al: **MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data.** *Genome Biol* 2015, **16:**278.
22. Reinius B, Sandberg R: **Random monoallelic expression of autosomal genes: stochastic transcription and allele-level regulation.** *Nat Rev Genet* 2015, **16:**653-664.
23. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, Lonnerberg P, Linnarsson S: **Quantitative single-cell RNA-seq with unique molecular identifiers.** *Nat Methods* 2014, **11:**163-166.
24. Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, Teichmann SA, Marioni JC, Stegle O: **Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells.** *Nat Biotechnol* 2015, **33:**155-160.
25. Risso D, Ngai J, Speed TP, Dudoit S: **Normalization of RNA-seq data using factor analysis of control genes or samples.** *Nat Biotechnol* 2014, **32:**896-902.
26. Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, Gingeras TR, Oliver B: **Synthetic spike-in standards for RNA-seq experiments.** *Genome Res* 2011, **21:**1543-1551.
27. Leek JT, Storey JD: **Capturing heterogeneity in gene expression studies by surrogate variable analysis.** *PLoS Genet* 2007, **3:**1724-1735.
28. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK: **Understanding mechanisms underlying human gene expression variation with RNA sequencing.** *Nature* 2010, **464:**768-772.
29. Stegle O, Parts L, Durbin R, Winn J: **A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies.** *PLoS Comput Biol* 2010, **6:**e1000770.
30. Kang HM, Ye C, Eskin E: **Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots.** *Genetics* 2008, **180:**1909-1925.
31. Brennecke P, Anders S, Kim JK, Kolodziejczyk AA, Zhang XW, Proserpio V, Baying B, Benes V, Teichmann SA, Marioni JC, Heisler MG: **Accounting for technical noise in single-cell RNA-seq experiments.** *Nature Methods* 2013, **10:**1093-1095.
32. Carvalho CM, Chang J, Lucas JE, Nevins JR, Wang QL, West M: **High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics.** *Journal of the American Statistical Association* 2008, **103:**1438-1456.
33. Pournara I, Wernisch L: **Factor analysis for gene regulatory networks and transcription factor activity profiles.** *BMC Bioinformatics* 2007, **8:**61.
34. Lucas JE, Kung HN, Chi JT: **Latent factor analysis to discover pathway-associated putative segmental aneuploidies in human cancers.** *PLoS Comput Biol* 2010, **6:**e1000920.
35. Blum Y, Le Mignon G, Lagarrigue S, Causeur D: **A factor model to analyze heterogeneity in gene expression.** *BMC Bioinformatics* 2010, **11:**368.
36. Parts L, Stegle O, Winn J, Durbin R: **Joint genetic analysis of gene expression data with inferred cellular phenotypes.** *PLoS Genet* 2011, **7:**e1001276.

37. Burns JC, Kelly MC, Hoa M, Morell RJ, Kelley MW: **Single-cell RNA-Seq resolves cellular complexity in sensory organs from the neonatal inner ear.** *Nat Commun* 2015, **6:**8557.
38. Tung J, Zhou X, Alberts SC, Stephens M, Gilad Y: **The genetic architecture of gene expression levels in wild baboons.** *Elife* 2015, **4**.
39. Lea AJ, Tung J, Zhou X: **A Flexible, Efficient Binomial Mixed Model for Identifying Differential DNA Methylation in Bisulfite Sequencing Data.** *PLoS Genet* 2015, **11:**e1005650.
40. West M: **Bayesian factor regression models in the "Large p, Small n" paradigm.** *Bayesian Statistics 7* 2003**:**733-742.
41. Law CW, Chen Y, Shi W, Smyth GK: **voom: Precision weights unlock linear model analysis tools for RNA-seq read counts.** *Genome Biol* 2014, **15:**R29.
42. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: **limma powers differential expression analyses for RNA-sequencing and microarray studies.** *Nucleic Acids Res* 2015, **43:**e47.
43. Soneson C, Delorenzi M: **A comparison of methods for differential expression analysis of RNA-seq data.** *BMC Bioinformatics* 2013, **14:**91.
44. Seyednasrollah F, Laiho A, Elo LL: **Comparison of software packages for detecting differential expression in RNA-seq studies.** *Brief Bioinform* 2015, **16:**59-70.
45. Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, Gonzalez-Porta M, Kurbatova N, Griebel T, Ferreira PG, et al: **Transcriptome and genome sequencing uncovers functional variation in humans.** *Nature* 2013, **501:**506-511.
46. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, Haudenschild CD, Beckman KB, Shi J, Mei R, et al: **Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals.** *Genome Res* 2014, **24:**14-24.
47. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET: **Transcriptome genetics using second generation sequencing in a Caucasian population.** *Nature* 2010, **464:**773-777.
48. Lee S, Chugh PE, Shen H, Eberle R, Dittmer DP: **Poisson factor models with applications to non-normalized microRNA profiling.** *Bioinformatics* 2013, **29:**1105-1111.
49. Zhou M, Hannah L, Dunson D, Carin L: **Beta-negative binomial process and Poisson factor analysis.** *Artificial Intelligence and Statistics* 2012, **22:**1462-1471.
50. Gene Ontology C: **The Gene Ontology project in 2008.** *Nucleic Acids Res* 2008, **36:**D440-444.
51. Santos A, Wernersson R, Jensen LJ: **Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes.** *Nucleic Acids Res* 2015, **43:**D1140-1144.
52. Zhou X, Stephens M: **Genome-wide efficient mixed-model analysis for association studies.** *Nat Genet* 2012, **44:**821-824.
53. Zhou X, Stephens M: **Efficient multivariate linear mixed model algorithms for genome-wide association studies.** *Nat Methods* 2014, **11:**407-409.

Figure legends

Figure 1: Illustration of scPLS. We model the expression level of genes in the control set (**X**) and genes in the target set (**Y**) jointly. Both control and target genes are affected by common confounding factors (**Z**) with effects $\mathbf{\Lambda}_x$ and $\mathbf{\Lambda}_y$ in the two sets, respectively. The target genes are also influenced by biological factors (**U**) with effects $\mathbf{\Lambda}_u$. The biological factors represent intermediate factors that coordinately regulate a set of genes, and are introduced to better capture the complex variance structure in the target genes. scPLS aims to remove the confounding effects $\mathbf{z}\mathbf{\Lambda}_y$ in the target genes.

Figure 2. Method comparison in simulations. (a) Compared with PCA (green) and LMM (purple), scPLS (blue) achieves higher power to identify DE genes across a range of effect sizes. The uncorrected data (orange) is also included in the panel. Power is evaluated at an empirical false discovery (FDR) rate of 0.05 in each of the 10 replicates. x-axis shows the effect sizes, which are measured as the percentage of phenotypic variation (PVE) in expression levels explained by the group label (ranges from 1% to 20%). (b) Sensitivity analysis shows that, compared with PCA and LMM, scPLS has the least reduction in power when a smaller subset of control genes are used (N=10, 20, 30 or 40 instead of 50). Power is averaged across replicates. (c) The first two principal components (PC) based on the top 100 identified DE genes from scPLS, but not from PCA or LMM or uncorrected data, cluster cells into two expected groups (red star and green triangle). (d) scPLS corrected expression data (blue) better classifies cells into the two known clusters than PCA (green) or LMM (purple) corrected expression data or uncorrected data (orange) from one replicate. Classification is based on support vector machine (SVM) with five-fold cross-validation. Accuracy is computed as the mean percentage of true positives in the test set across replicates.

Figure 3. Removing technical effects in the Grun et al. data. (a) For each gene, the expression variance is partition into three components: a component that is explained by technical confounding factors (blue), a component that is explained by biological factors (orange), and the residual error variance (purple). Genes are evenly divided into ten quantiles based on the sample variance. (b) Heat map visualizes gene expression levels in the uncorrected data and scPLS corrected data. The plot reveals reduction in within group variation after correction. (c) PCs from scPLS corrected data (blue) can better reveal the known cells clusters than that from the PCA (purple), LMM (orange), or RUV (green) corrected data. For each PC, we evaluate its ability to classify cells into the correct clusters before and after correction. We compute the reduction in within-cluster variance as a measure of the improvement after correction; a higher value thus indicates better performance.

Figure 4. Removing cell cycle effects in the Burns et al. data. (a) For each gene, the expression variance is partition into three components: a component that is explained by cell cycle factors (blue), a component that is explained by biological factors (orange), and the residual error variance (purple). Genes are evenly divided into ten quantiles based on the sample variance. (b) PCA plot for the uncorrected data and scPLS corrected data. Six different cell types are in different color and shape. Notice the similarity between PC3 vs PC2 plot in the uncorrected data and PC2 vs PC1 plot in the scPLS corrected data. (c) Correlation of the first three PCs from either the uncorrected data (orange) or data corrected by different methods with the first PC of cell cycle genes (ccPC). Correlation is displayed on log scale. (d) Heatmap of correlations between control genes and validation genes in the subset analysis. In uncorrected data, genes from the two sets are intermingled. After scPLC correction, the two sets are easily separable. (e) In the subset analysis, pairwise correlations are computed between genes from the controls set and that from the validation set. The median correlation, as well as correlations at the up quantiles from scPLS are lower than the other methods.

Figure 5. Removing both technical effects and cell cycle effects in the Buettner et al. data. (a) For each gene, the expression variance is partition into four components: a component that is explained by technical confounding factors (blue), a component that is explained by cell cycle factors (green), a component that is explained by biological factors (orange), and the residual error variance (purple). Genes are evenly divided into ten quantiles based on the sample variance. (b) PCA analysis for the uncorrected data and scPLS corrected data. In the uncorrected data, there is a clear separation of cells by cell-cycle stage. Such separation of cells is no longer observed in the scPLS corrected data. (c) Estimated proportion of variance explained by the cell cycle inferred from different methods are plotted against that obtained with the known cell cycle labeling. scPLS achieves a higher correlation than other methods.

Supplementary Figure legends

Figure S1. Sensitivity analysis compares the estimation accuracy of scPLS with that of PCA and LMM in simulations. The three methods are applied to remove confounding effects using either 10, 30 or 50 control genes. For each gene, the estimated proportion of variance explained by the confounding effects is plotted against the truth. scPLS achieves higher correlation between the estimates and the truth. In addition, scPLS is less sensitive to the number of control genes used in the analysis.

Figure S2. Power comparison in simulations where the confounding factors are not corrected with the group label. Compared with PCA (green) and LMM (purple), scPLS (blue) achieves higher power to identify DE genes across a range of effect sizes. The uncorrected data (orange) is also included in the panel. Power is evaluated at an empirical false discovery (FDR) rate of 0.05 and averaged across 10 replicates. The effect sizes are measured as the percentage of phenotypic variation (PVE) in expression levels explained by the group label (ranging from 1% to 20%).

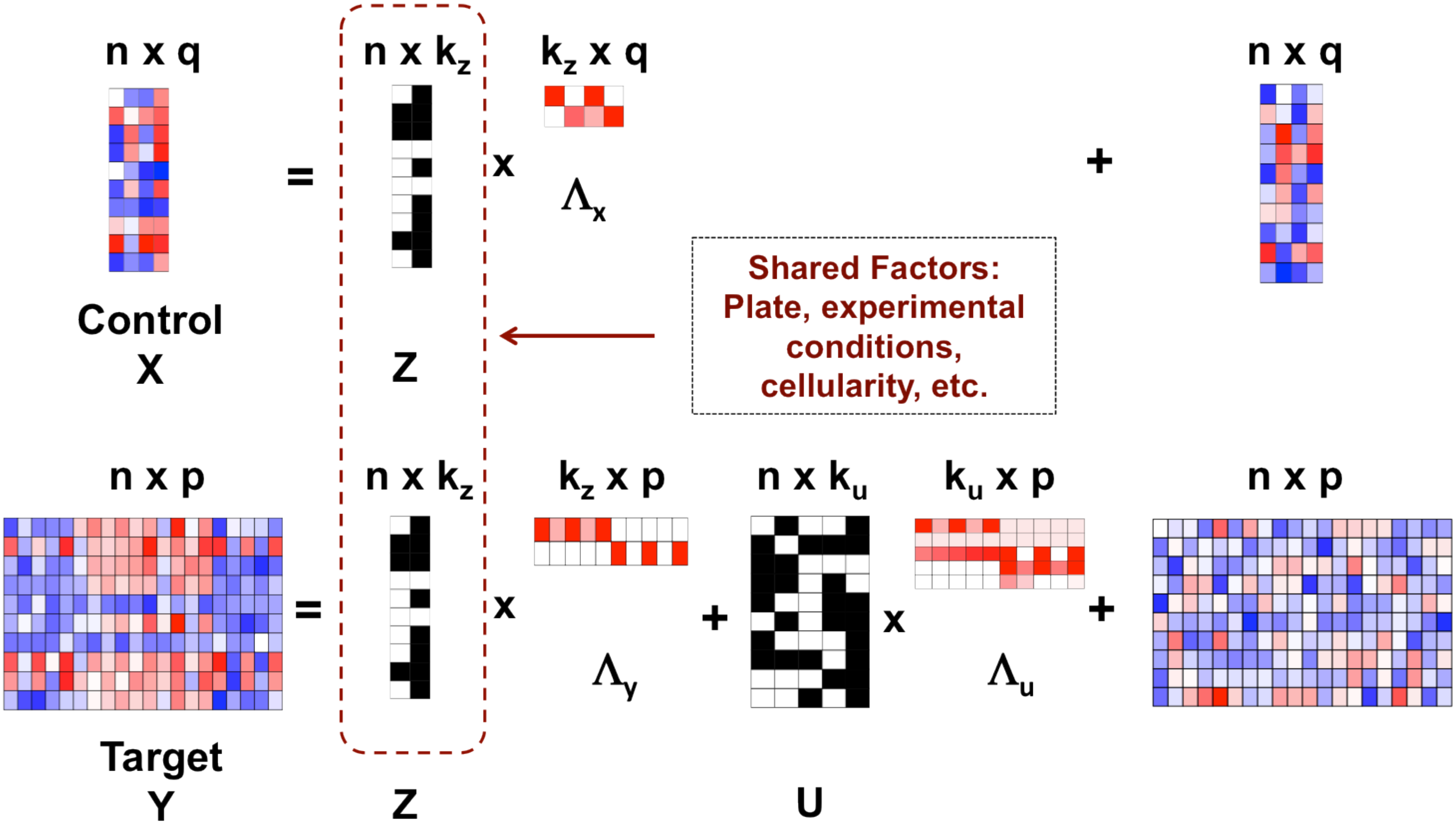
Figure S3. (a) Power comparison of the full model (scPLS) and the reduced model (reduced scPLS) in the presence of biological factors. PCA, LMM and uncorrected data are also included. Power is evaluated at an empirical false discovery (FDR) rate of 0.05 across 10 replicates. x-axis shows the effect sizes, which are measured as the percentage of phenotypic variation (PVE) in expression levels explained by the group label (ranges from 1% to 20%). (b) Power comparison of full model using different number of biological factors (K=2, 4, 6, 8) and the reduced model (K=0) when there is no biological factors in the data. Power is evaluated at an empirical false discovery (FDR) rate of 0.05 across 10 replicates. Under this setting, the full model does not perform much worse that the reduced model with respect to the power of identifying DE genes. The comparison results are also not sensitive with respect to the number of factors used in the model even when the truth is zero.

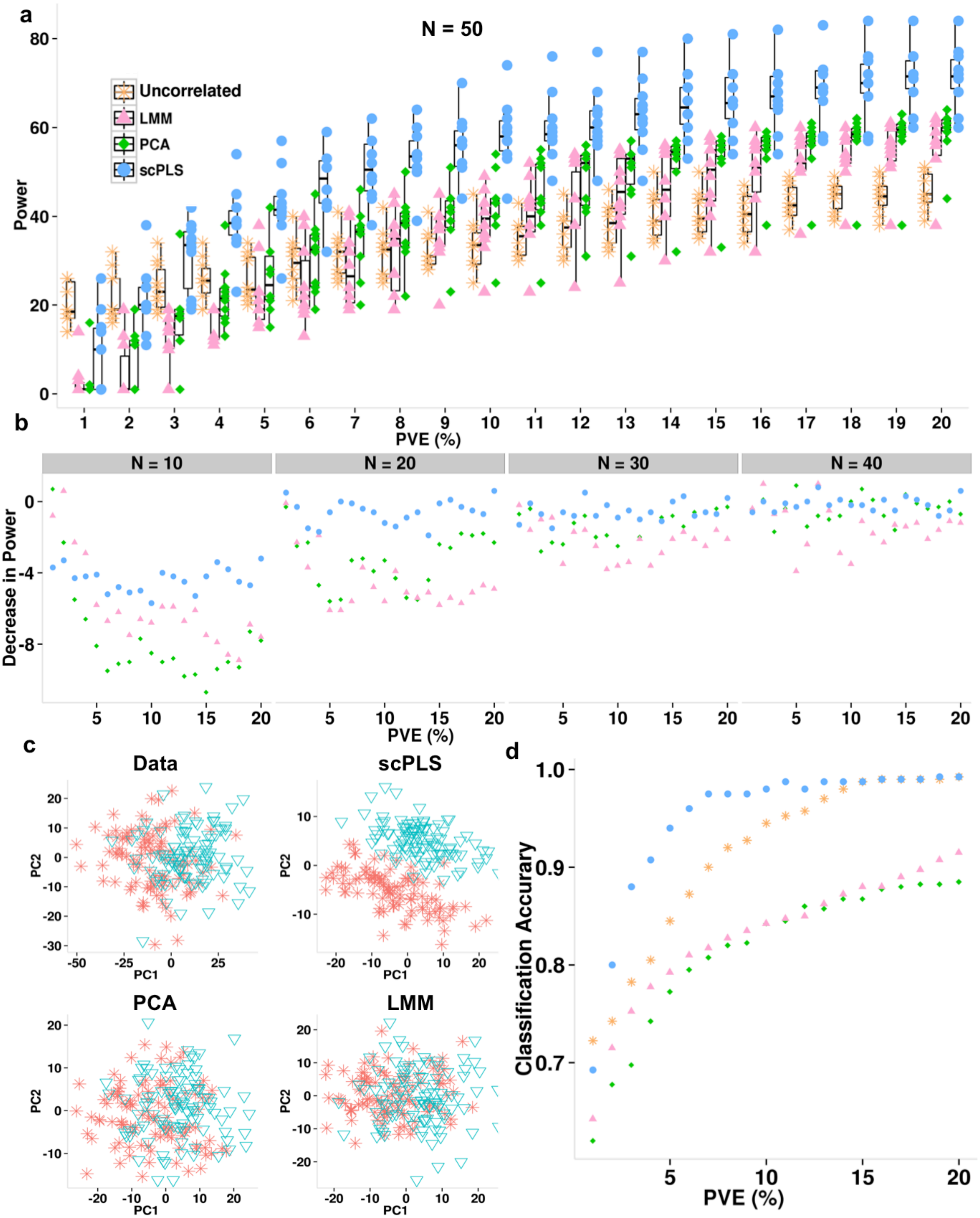
Figure S4. The comparison of estimates of cell cycle proportion for genes in the mouse embryonic stem cell data using different scPLS algorithms: EM-in-chunks (as in the paper), naïve EM and EM-in-chunks using a subset of cell cycle genes. We only perform comparison in this data because we know the true cell cycle labels.

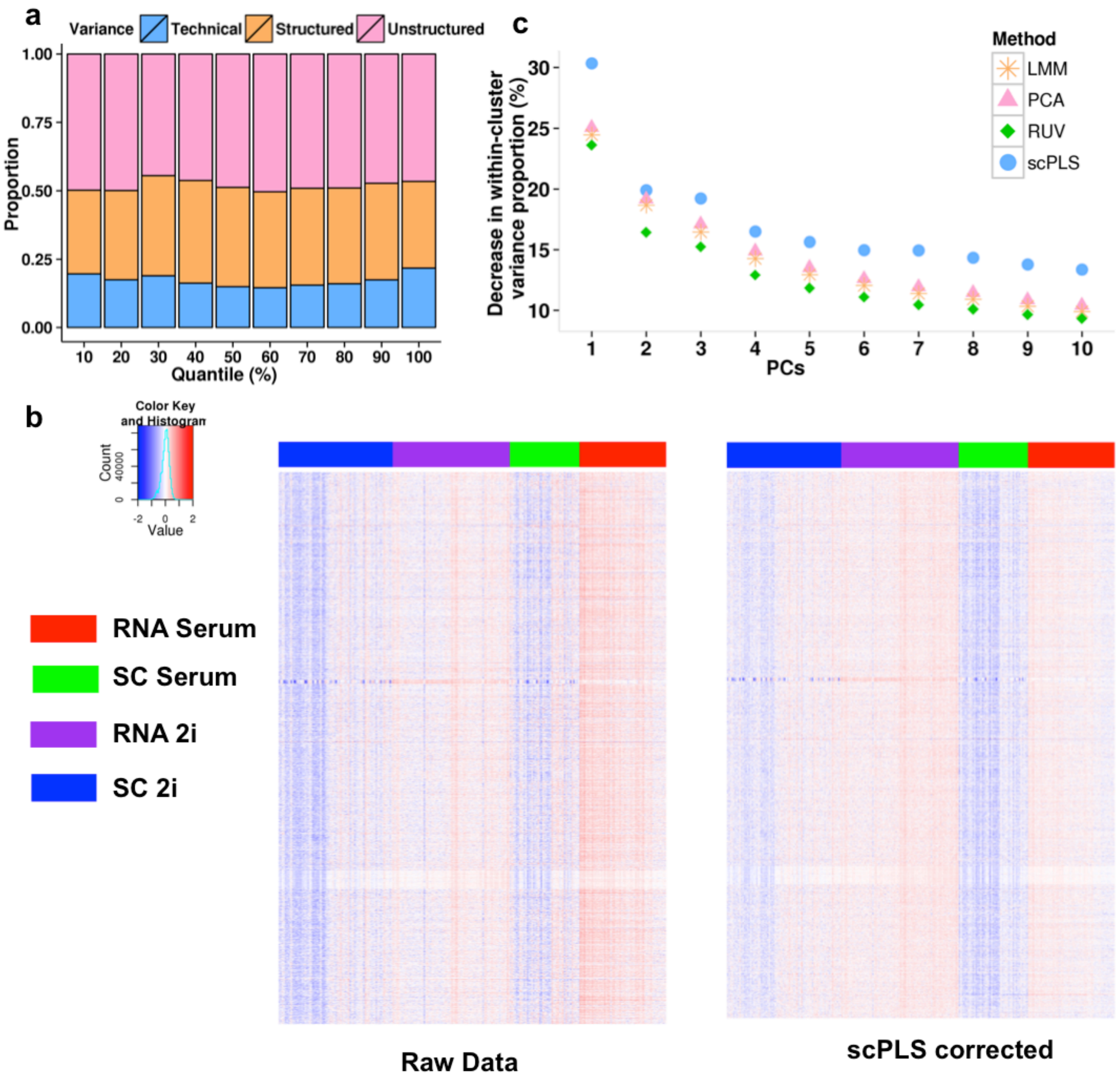
Figure S5. (a) In the subset analysis of the Buettner et al. data, pairwise correlations are computed between genes from the controls set and that from the validation set. The median correlation, as well as correlations at the up quantiles from scPLS are lower than the other methods. (b) Heatmap of correlations between control genes and validation genes in the subset analysis of the Buettner et al. data. In uncorrected data, genes from the two sets are intermingled. After scPLC correction, the two sets are easily separable.

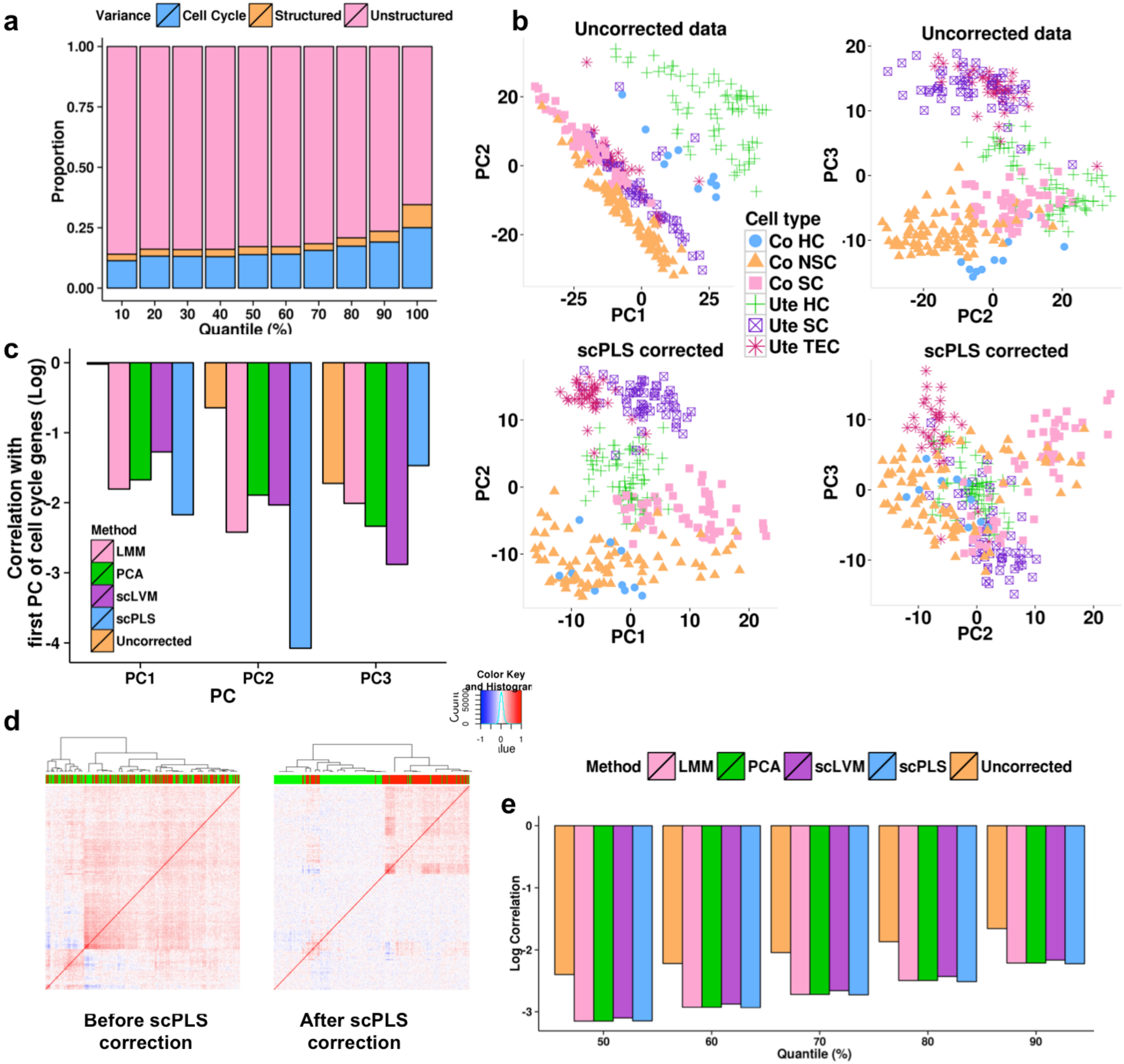
Table S1. Comparison of the naïve EM algorithm and the EM-in-chunks algorithm in terms of accuracy and speed in simulations. The EM-in-chunks algorithm uses either a chunk size of 500 genes or a chunk size of 1,000 genes. Accuracy is measured by the estimation errors of the loading matrix in terms of the normalized Frobenius norm. Speed is measured by CPU time in seconds. n: the number of cells. p: the number of genes.

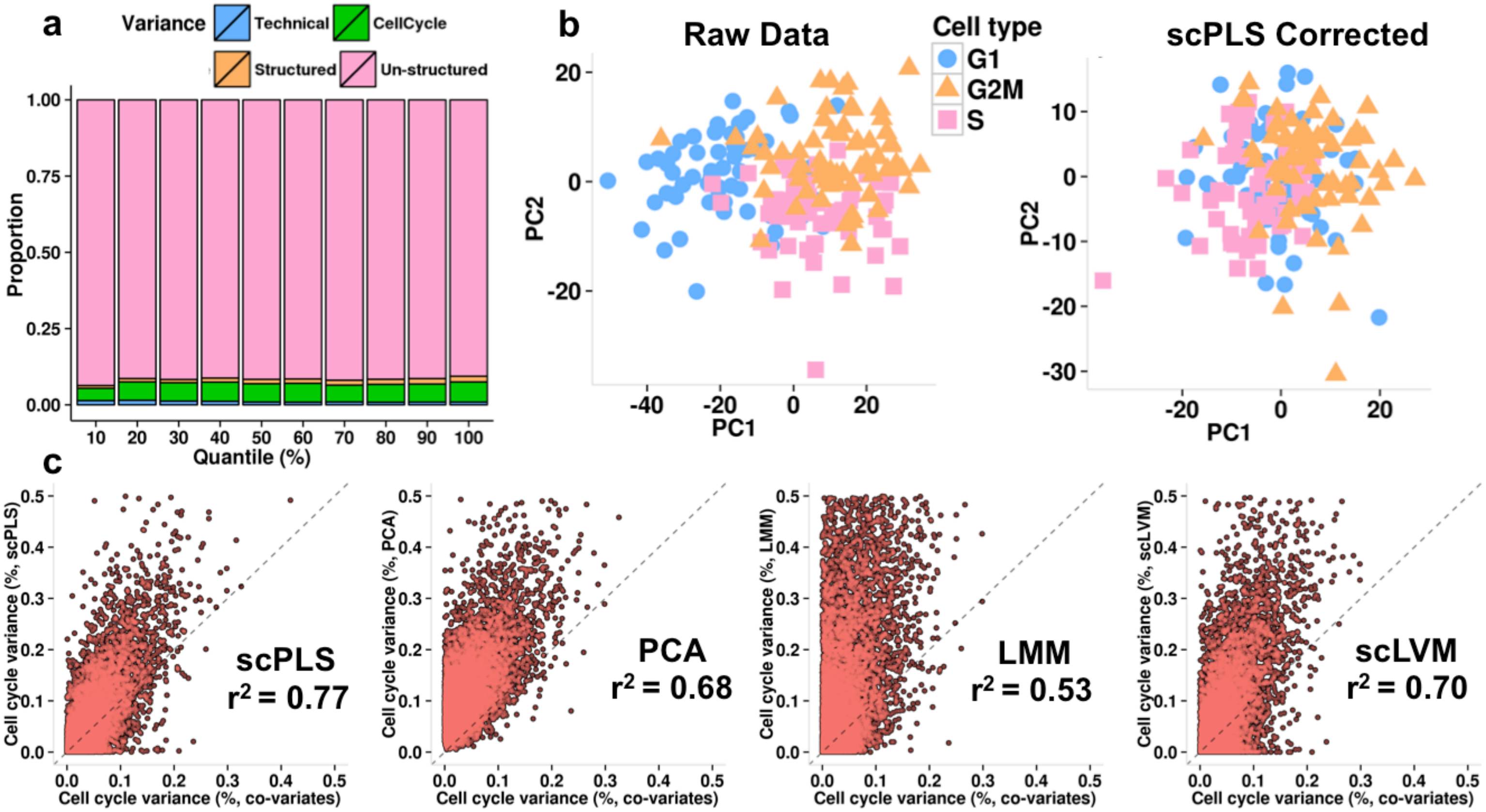
Supplementary File 1 This zip files contain gene expression levels of three described datasets before and after correcting for technical or cell cycle effect.


n x q
Control
X
=
n x k_z
Z
x
k_z x q
Λ_x
+
n x q
Shared Factors:
Plate, experimental
conditions,
cellularity, etc.
n x p
Target
Y
=
n x k_z
Z
x
k_z x p
Λ_y
+
n x k_u
U
x
k_u x p
Λ_u
+
n x p


a
N = 50
Uncorrelated
LMM
PCA
scPLS
Power
80
60
40
20
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
PVE (%)
b
N = 10
N = 20
N = 30
N = 40
Decrease in Power
0
-4
-8
5
10
15
20
PVE (%)
c
Data
scPLS
PCA
LMM
PC1
PC2
-50
-25
0
25
-30
-20
-10
0
10
20
d
Classification Accurary
1.0
0.9
0.8
0.7
5
10
15
20
PVE (%)


a
Variance
Technical
Structured
Unstructured
Proportion
1.00
0.75
0.50
0.25
0.00
10
20
30
40
50
60
70
80
90
100
Quantile (%)
c
Method
LMM
PCA
RUV
scPLS
Decrease in within-cluster variance proportion (%)
30
25
20
15
10
1
2
3
4
5
6
7
8
9
10
PCs
b
Color Key and Histogram
Count
0
40000
-2
0
2
Value
RNA Serum
SC Serum
RNA 2i
SC 2i
Raw Data
scPLS corrected


a
Variance
Cell Cycle
Structured
Unstructured
Proportion
1.00
0.75
0.50
0.25
0.00
10
20
30
40
50
60
70
80
90
100
Quantile (%)
b
Uncorrected data
PC2
PC1
20
0
-20
-25
0
25
Uncorrected data
PC3
PC2
20
10
0
-10
-20
0
20
Cell type
Co HC
Co NSC
Co SC
Ute HC
Ute SC
Ute TEC
scPLS corrected
PC2
PC1
10
0
-10
-10
0
10
20
scPLS corrected
PC3
PC2
10
0
-10
-10
0
10
20
c
Correlation with first PC of cell cycle genes (Log)
0
-1
-2
-3
-4
Method
LMM
PCA
scLVM
scPLS
Uncorrected
PC1
PC2
PC3
PC
Color Key and Histogram
Count
0
50000
-1
0
1
Value
d
Before scPLS correction
After scPLS correction
Method
LMM
PCA
scLVM
scPLS
Uncorrected
e
Log Correlation
0
-1
-2
-3
50
60
70
80
90
Quantile (%)


a
Variance
Technical
CellCycle
Structured
Un-structured
Proportion
1.00
0.75
0.50
0.25
0.00
10
20
30
40
50
60
70
80
90
100
Quantile (%)
b
Raw Data
Cell type
G1
G2M
S
scPLS Corrected
PC2
PC1
20
0
-20
-40
10
-10
-30
c
scPLS
r² = 0.77
PCA
r² = 0.68
LMM
r² = 0.53
scLVM
r² = 0.70
Cell cycle variance (%, scPLS)
Cell cycle variance (%, PCA)
Cell cycle variance (%, LMM)
Cell cycle variance (%, scLVM)
Cell cycle variance (%, co-variates)
0.0
0.1
0.2
0.3
0.4
0.5