A robust statistical framework to detect multiple sources of hidden variation in single-cell transcriptomes

Donghyung Lee[1,*], Anthony Cheng[1,2], and Duygu Ucar[1,*]

[1]The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut, Unites States of America, [2]University of Connecticut Health Center, Farmington, Connecticut, Unites States of America

*Correspondence: *donghyung.lee@jax.org* and *duygu.ucar@jax.org*

Abstract

Single-cell RNA-Sequencing data often harbor variation from multiple correlated sources, which cannot be accurately detected by existing methods. Here we present a novel and robust statistical framework that can capture correlated sources of variation in an iterative fashion: iteratively adjusted surrogate variable analysis (IA-SVA). We demonstrate that IA-SVA accurately captures hidden variation in single cell RNA-Sequencing data arising from cell contamination, cell-cycle stage, and differences in cell types along with the marker genes associated with the source.

Single-cell RNA-Sequencing (scRNA-Seq) data often harbor variation from diverse sources including technical (e.g., biases in capturing transcripts from single cells, PCR amplifications) and biological factors (e.g., differences in cell cycle stage or cell types) that might confound biological conclusions [1-3]. Detecting and adjusting for hidden heterogeneity in scRNA-Seq data is essential to accurately characterize gene expression changes stemming from a biological variable of interest (e.g., disease vs. normal). A number of statistical methods have been proposed to detect hidden sources of variation in microarray, bulk, and single-cell RNA-Seq data: SSVA[4] (supervised surrogate variable analysis), USVA[5] (unsupervised SVA), ISVA[6] (Independent SVA), RUVcp[7, 8] (removing unwanted variation using control probes), RUVres (RUV using residuals), RUVemp (RUV using empirical negative controls) and scLVM[9] (single-cell latent variable model). One caveat of these methods is their assumption that the multiple sources of variation are uncorrelated (i.e., orthogonal) with each other and with known variables[6]. However, in reality transcriptomic data especially single cell measurements typically contain variation stemming from multiple yet correlated hidden factors due to poor experimental design, technical limitations, or biological factors. For example, the number of expressed genes in a cell (a major source of variation), experimental batch effects, cell cycle stage, cell size, and cell type can be highly correlated with each other and may confound the downstream biological conclusions [9, 10 11, 12]. To properly detect and account for these sources of variation, we developed a robust and iterative statistical framework, IA-SVA (iteratively adjusted surrogate variable analysis) (**Fig. 1a**). IA-SVA is designed to identify multiple and potentially correlated hidden sources of variation from scRNA-Seq data with high statistical power and low error rate (see Online Methods, **Supplementary Fig. 1**, and https://github.com/UcarLab/IA-SVA/).

The major advantages of IA-SVA over existing methods are three-fold: First, it accurately captures multiple hidden sources of variation even if the sources are correlated. Second, it enables assessing the significance of each detected factor for explaining the unmodeled variation in the data. Third, it delivers marker genes that are significantly associated with the detected hidden factors. Factors or marker genes inferred by IA-SVA can be instrumental in data interpretation and in improving the performance of downstream analyses, such as clustering/visualization of single-cell data using t-distributed stochastic neighbor embedding (t-SNE) [13].

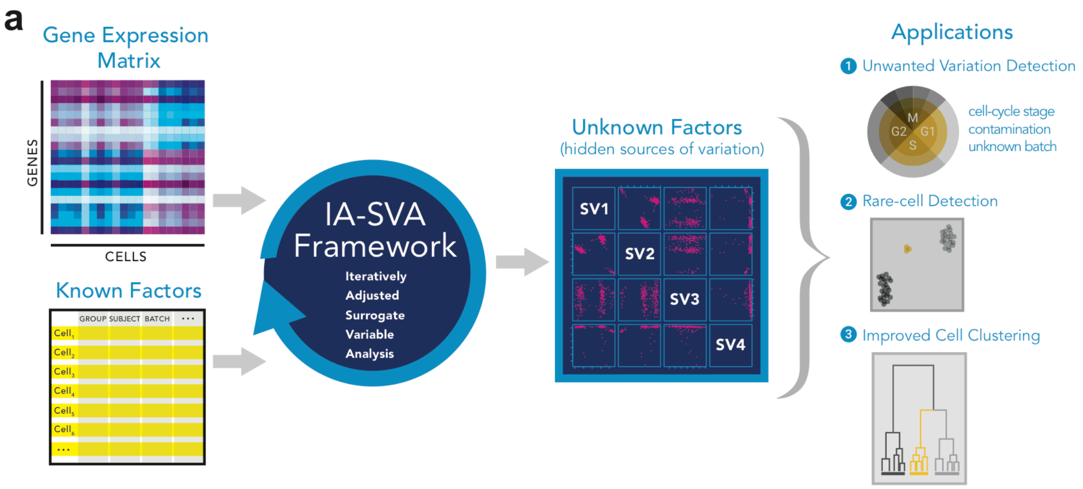
Using simulated scRNA-Seq data, we studied and compared the empirical Type I error rate, the power of detection, and the accuracy of estimation for IA-SVA and existing state-of-the-art methods, which can also infer the number of significant hidden factors (i.e., USVA and SSVA) (See Online Methods). Under different simulation scenarios, we found that IA-SVA consistently outperformed USVA and SSVA in terms of detection power and accuracy of the estimate while controlling the Type I error rate under the nominal level (0.05) (**Fig. 1b**). In particular, IA-SVA significantly outperformed alternatives when hidden factors affect a small percentage of genes (10-20%) and when these factors are moderately correlated with a known factor (i.e., group variable) (the first three columns of **Fig. 1b**). We compared the efficacy of IA-SVA against a broader number of supervised (SSVA and RUVcp) and unsupervised (USVA, PCA, RUVemp and RUVres) methods (**Supplementary Note 1)**. Similarly, IA-SVA was particularly effective in estimating hidden factors that affect a subset of genes (10-20%) (Factor 3 in **Supplementary Fig. 2)** and in inferring correlations among factors (**Supplementary Fig. 3**). We also compared the performance of IA-SVA against unsupervised methods (USVA, PCA, RUVemp, RUVres) to estimate the heterogeneity arising from differences in brain cell types (neurons vs. oligodendrocytes) [14] (See Online Methods). IA-SVA significantly outperformed other methods and accurately inferred the factor that corresponds to cell type assignments ($|r|$ = 0.95 vs. 0.83 for the second best performance by RUVres) (**Supplementary Fig. 4**).

To test the efficacy of IA-SVA in capturing variation within a relatively homogenous cell population, we studied alpha cells (n=101) from three diabetic patients [15] (see Online Methods). We found that Surrogate Variable 2 (SV2) inferred by IA-SVA clearly separated alpha cells into two groups (six outlier cells marked in red vs. the rest at SV2 < -0.2) (**Fig. 2a**). Top 30 genes (e.g., *CD9*, *SPARC, COL4A1, PMEPA1, ENG*) correlated with SV2 clearly separated alpha cells into two clusters, where six outlier cells exclusively expressed these genes (**Fig. 2b**). Alternative methods (PCA, USVA, tSNE) didn't clearly separate these outlier cells, especially in the case of tSNE analyses (**Fig. 2a**). This heterogeneity detected in alpha cells was reproducible in a bigger and independently generated islet scRNA-Seq data using the same platform [16] (**Supplementary Fig. 5**). In both datasets this heterogeneity was associated with fibrotic response genes (e.g., *SPARC, COL4A1, COL4A2)* suggesting that these outlier cells might originate from cell contamination (e.g., fibroblasts contaminating islet cells) or from cell doublets captured together—a known problem in early Fluidigm C1 experiments [17, 18].

Another established source of heterogeneity in scRNA-Seq data is the differences in cell-cycle stages [3]. To test whether IA-SVA can capture this, we analyzed scRNA-seq data obtained from human glioblastomas with an established cell-cycle signature [19]. Using IA-SVA, we detected a source of hidden heterogeneity (SV2) that clearly separated 12 cells from the rest (**Fig. 2c**) and identified 87 marker genes associated with this source (**Fig. 2d**). Pathway and GO enrichment analyses of these marker genes [20, 21] revealed significant enrichment for cell-cycle stage related GO terms and KEGG pathways (**Supplementary Fig. 6 and Supplementary Table 1**). PCA, USVA and tSNE failed to separate these cells (**Fig. 2c**).

In scRNA-Seq data, technical or biological factors are often correlated and can deteriorate the single cell clustering results (e.g., clustering with respect to cell types) by masking the real signal or generating spurious clusters. IA-SVA can be particularly effective in handling this problem by uncovering hidden factors while adjusting for all potential confounders. Moreover, IA-SVA delivers marker genes associated with the hidden factor, which can be further tested and evaluated for their biological relevance (e.g., novel markers for different cell types) and can be utilized in clustering analyses for increased performance. To test this, we first studied scRNA-Seq data from alpha (n=101), beta (n=96), and ductal (n=16) cells obtained from three diabetic patients [15] (Online Methods) and used tSNE on all expressed genes to cluster these cells. Color-coding based on the reported cell type assignments [15] showed that, tSNE cannot effectively separate these cells into their respective categories (**Fig. 3a**). Next, we applied IA-SVA on this data and focused on top two significant SVs (SV1 and SV2) since they separated cells into distinct clusters (**Supplementary Fig. 7**). 86 genes were associated with these two SV2 that notably included previously known markers used in the original study (*INS*, *GCG*, *KRT19*) and uncovered potential novel markers of islet cells (Fig. 3c). As expected, tSNE analyses on these 86 genes improved the clustering results significantly and clearly separated different cell types (**Fig. 3b**). Such improved clustering analyses can also help reveal cells that might be incorrectly labeled based on a single gene marker. We tested whether this pattern can be recapitulated in a bigger data with confounding variables[16] by analyzing transcriptomes of 1600 islet cells including alpha (n=946), beta (n=503), delta (n=58), and PP (n=93) cells (Online Methods). In this case, designated cell type assignments correlated with known factors especially with the patient identifications (*C*=0.48 for patient id, *C*=0.1 for sex, *C*=0.03 for phenotype and *C*=0.25 for ethnicity, *C*=Pearson's contingency coefficient). If not properly adjusted for, these correlations would lead to spurious clustering of cells. For example, when tSNE is performed on these islet cells and cells are color-coded with respect to the original cell-type assignments [16], cell types did not separate from each other and spurious clusters were observed within each cell type (**Fig. 3d**). As suspected, potential confounding factors, particularly patient id and ethnicity, explained the spurious clustering of cells (**Supplementary Fig. 8**). Existing methods to improve scRNA-Seq clustering results (e.g., 'Spectral tSNE' [22]) regress out (remove) variation associated with known variables before estimating hidden factors. However, when biological variables of interest (e.g., cell type assignments) are highly correlated with known factors as in this case, removing the known effects will also impact the signal of interest. To handle this, we conducted IA-SVA analyses while accounting for known factors and extracted four significant SVs. Among these, SV1 and SV4 grouped cells into disjoint clusters (**Supplementary Fig. 9a and b**); therefore we focused on these as putative SVs associated with differences in cell types (SV3 is not considered since it captures cell contamination). 57 genes associated with these two SVs included once again known marker genes for islet cells (i.e., *INS* and *GCG*) (**Supplementary Fig. 10**). tSNE analyses using these genes clearly separated different cell types into discrete clusters and reinforced the importance of properly adjusting for known factors prior to clustering or marker gene detection (**Figure 3e**). Top surrogate factors obtained via PCA and USVA failed to detect the heterogeneity associated with cell types (**Supplementary Fig. 9c and d**).

In summary, IA-SVA can accurately and robustly estimate hidden sources of variation in gene expression data while adjusting for known factors introducing unwanted variation. The iterative framework to detect multiple and potentially correlated factors along with their significance is the main advantage of IA-SVA over existing methods. This flexibility is more realistic given the confounded nature of known and unknown factors introducing heterogeneity in gene expression levels particularly in scRNA-Seq data. Furthermore, IA-SVA infers marker genes associated with the source of variation that can be used for various purposes including novel marker gene detection for different cell types.



b

	USVA	SSVA	IA-SVA	USVA	SSVA	IA-SVA
	|r| = 0.3 ~ 0.6			|r|<0.3		
Power*(F1**)	1	1	1	1	1	1
Power (F2)	1	1	1	1	1	1
Power (F3)	0.78	0.78	0.87	1	1	1
Cor***(F1)	0.93	0.95	0.95	0.98	0.98	1
Cor (F2)	0.72	0.75	0.94	0.94	0.94	0.99
Cor (F3)	0.75	0.78	0.95	0.93	0.93	0.98

	USVA	SSVA	IA-SVA
Type I error*	0.09	0.09	0.04

* Nominal Type I error rate: 0.05
** F1, F2, F3 refers to Factor1, Factor2, and Factor 3
*** Average of the absolute Pearson correlation coefficient between the true factor and the estimated factor is used as the accuracy measure.

Figure 1. IA-SVA is a robust statistical framework to detect sources of hidden heterogeneity. (a) IA-SVA uses single-cell gene expression data matrix and known factors to detect hidden sources of variation (e.g., cell contamination, cell-cycle status, and cell type). These hidden factors can be used as additional covariates in differential analysis to increase statistical power. If these factors match to a biological variable of interest (e.g., cell type assignment), genes highly correlated with the factor can be detected and used in downstream analyses (e.g., clustering). **(b)** Empirical Type I error rate, detection power and the accuracy of estimates for IA-SVA, SSVA, and USVA using simulated single-cell gene expression data. Alternative scenarios are simulated in which hidden factors are moderately (|*r*|=~0.3-0.6, first three columns) or weakly (|*r*|<0.3, last three columns) correlated with the group variable.

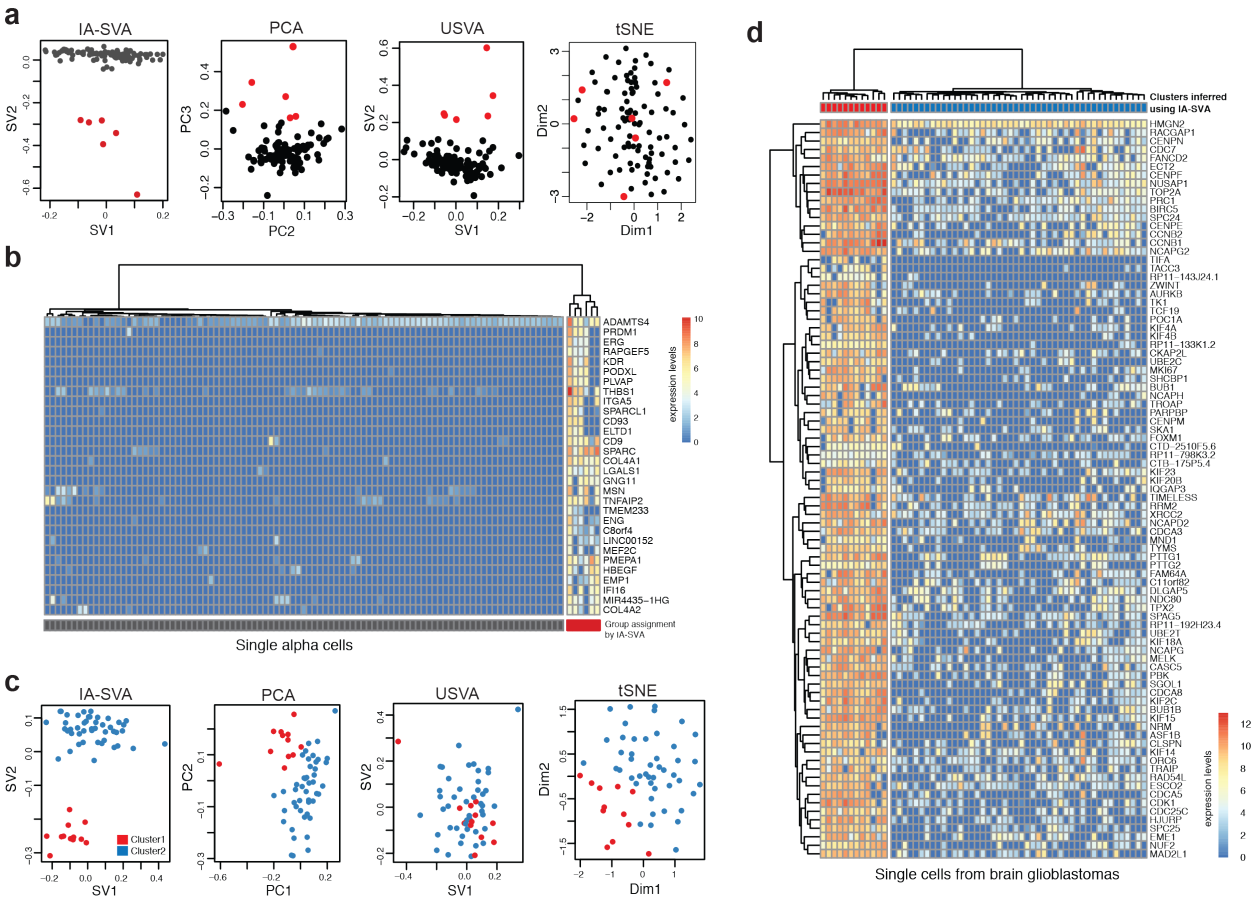


Figure 2. IA-SVA can detect heterogeneity originated from a few cells. (a) Heterogeneity within alpha cells captured using IA-SVA, PCA, USVA, and tSNE. Cells are clustered into two groups (black vs. red dots) based on IA-SVA's surrogate variable 2 (SV2 < -0.2). In PCA, PC1 was discarded since it explains the number of expressed genes. **(b)** Hierarchical clustering of alpha cells using the top 30 marker genes (ward.D2 and cutree_cols =2). 6 cells clearly separated from the rest of the cells in terms of the expression of these 30 genes. (**c)** Heterogeneity detected within glioblastomas using IA-SVA, PCA, USVA, and tSNE. IA-SVA's SV2 clearly separates cells into two groups (blue vs. red dots, SV2 < -0.1) with respect to their cell cycle stages. Other methods failed to detect this cell-cycle related heterogeneity. **(d)** Hierarchical clustering on 87 marker genes confirms the separation of cells based on these markers (ward.D2 and cutree_cols = 2).

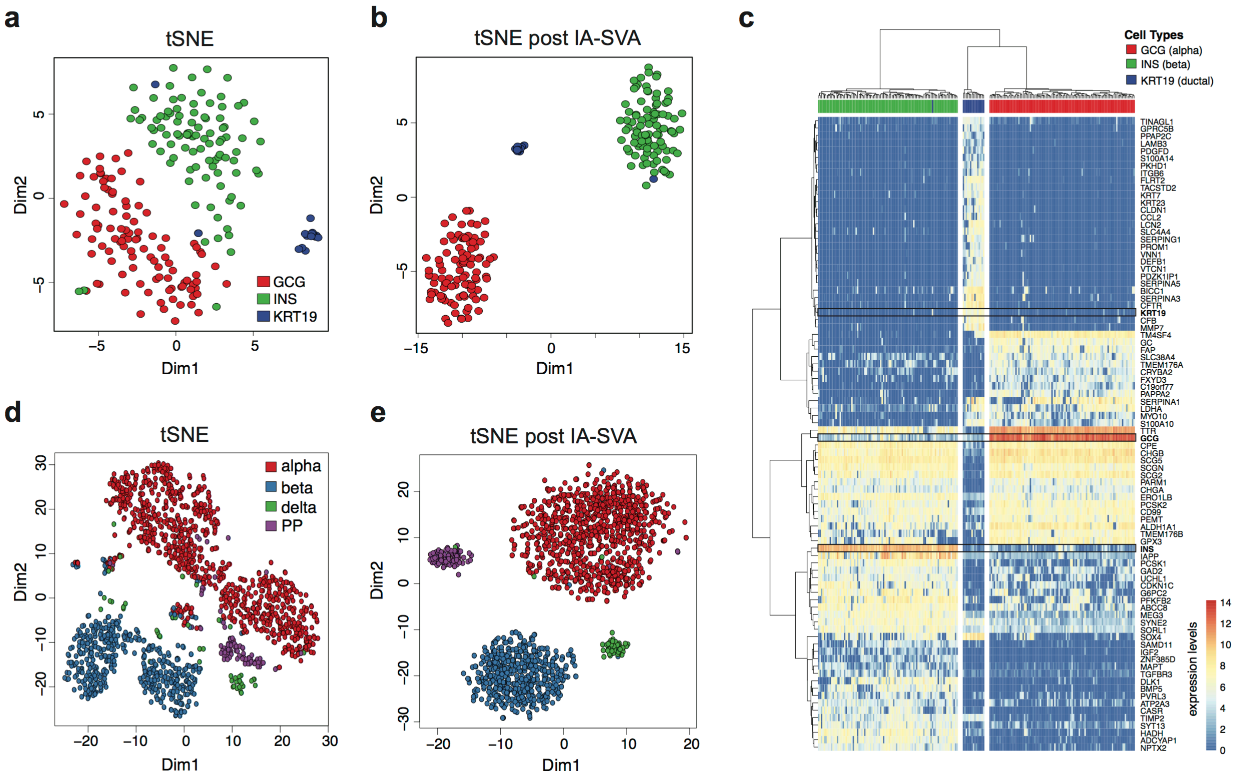


Fig. 3. IA-SVA based marker gene selection enhances the performance of clustering algorithms. (a) tSNE analyses using all expressed genes in human islet data. Cells are color-coded based on original cell-type assignments. **(b)** tSNE analyses using IA-SVA marker genes (n=86). Note the improved clustering of cell types into discrete clusters. **(c)** Hierarchical clustering using 86 marker genes clearly separate cell types (ward.D2 and cutree_cols=3). Rows marked with boxes refer to marker genes used in the original study. **(d)** tSNE analyses using all expressed genes in a bigger islet data. Note that cells are not effectively clustered with respect to their assigned cell types. **(e)** tSNE analyses using marker genes obtained via IA-SVA (n=57). Note the improved clustering of cells into discrete clusters.

ACCESSION CODES

The single-cell RNA sequencing read counts and annotations describing samples and experiment settings are included in an R data package ("iasvaExamples") containing data examples for IA-SVA (https://github.com/dleelab/iasvaExamples).

ACKNOWLEDGMENTS

This work has been supported by the Jackson Laboratory (JAX) for Genomic Medicine start-up funds (to D.U.) and the Jackson Laboratory Scientific Services Innovation Fund (to D.L. and D.U.). We thank JAX Computational Science group, Ucar and Stitzel lab members for constructive feedback throughout this project. We thank Jane Cha, JAX scientific illustrator, for her help with the figures.

AUTHOR CONTRIBUTIONS

D.L. and D.U. designed the project, generated the figures and wrote the manuscript. D.L. developed the statistical framework and run the data analyses. A.C. contributed to the data pre-processing and the generation of the R package. All authors read and approved this manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

ONLINE METHODS

IA-SVA framework. Formally, we model the log-transformed sequencing read counts for *m* genes and *n* samples (*i.e.,* $m \times n = Y$) as a combination of primary variable of interest, known and unknown sources of variation as follows:

$$Y_{m\times n} = X_{m\times p}\beta_{p\times n} + Z_{m\times q}\gamma_{q\times n} + W_{m\times k}\delta_{k\times n} + \varepsilon_{m\times n},$$

where X is a matrix for p primary variable(s) of interest (e.g., group assignment for cases and controls), Z is a matrix for q known factors (e.g., sex or ethnicity), W is a matrix for k unknown factors and ε is the error term. With this model, we can account for any clinical/experimental information about samples (e.g., sex, ethnicity, age, BMI, experimental batch) as known factors (Z) and dissect the variation in the read count data that is attributable to hidden factors (W).

Existing unsupervised methods (e.g., USVA, RUVres, ISVA) obtain the residual matrix by regressing read counts (Y) on all known factors (X and Z). Then, they infer the number of hidden factors and directly estimate hidden factors from the residual matrix using dimensionality reduction algorithms (e.g., principal component analysis (PCA), singular value decomposition (SVD) or independent component analysis (ICA)) under the assumption that hidden factors are uncorrelated with each other and also with the known factors. Consequently, when this assumption is not met, the direct inference from the residual matrix can lead to biased estimates of hidden factors and distort estimates.

In contrast, IA-SVA does not impose the assumption of uncorrelated factors. Instead, it allows correlations between factors to accurately estimate hidden factors via a novel iterative approach. At each iteration, IA-SVA obtains residuals, i.e., read counts adjusted for all known factors (X and Z) including unknown factors (surrogate variables) estimated from previous iterations and extracts the principal component (PC1) from the residuals using SVD. Next it tests the significance of PC1 in terms of its contribution to the unmodeled variation (i.e., the variation of residuals). Using this PC1 (as in the case of previous methods) as a surrogate variable assumes known factors and hidden factors are not correlated. Therefore, IA-SVA uses PC1 to infer marker genes associated with the hidden factor by taking advantage of the fact that PC1 and the true hidden factor are highly correlated. To detect these marker genes, IA-SVA regresses *Y* on PC1 and calculates the coefficient of determination (R^2) for each gene. Genes with high R^2 scores are considered as marker genes associated with the hidden factor. These genes are used for an unbiased inference of the hidden factor. For this, IA-SVA weighs all genes with respect to their R^2 scores, conducts SVD on the weighted read count matrix to obtain an unbiased PC1, and use this PC1 as a surrogate variable (SV) for the hidden factor. In the next iteration, IA-SVA uses this SV as an additional known factor to identify further significant hidden factors. The iterative procedure of IA-SVA composed of six major steps as summarized in **Supplementary Figure 1** and below:

[Step 1] Regress *Y* on all known factors (*X* and *Z)*, including a surrogate variable (SV) obtained from the previous iteration, to obtain residuals.

[Step 2] Conduct SVD on the obtained residuals to extract the first PC (PC1).

[Step 3] Test the significance of the contribution of PC1 to unexplained variation in the read count matrix *(Y)* using a non-parametric permutation-based assessment [5, 23, 24]. For more details, see next section.

[Step 4] If PC1 is significant, regress *Y* (in this case not using the known variables) on PC1 to compute the coefficient of determination (R^2) for every gene. If PC1 is not significant, stop the iteration and conduct subsequent down stream analysis using previously obtained significant

SVs.

[Step 5] Weigh each gene in *Y* with respect to its R^2 value by multiplying a gene's read counts with its R^2 values. The highly weighted genes in this framework serve as the marker genes for the hidden factor.

[Step 6] Conduct a second SVD on this weighted *Y* to obtain the first PC, which will be used as the surrogate variable (SV) for the hidden factor.

At the end of this six-step procedure, if a significant SV is obtained, IA-SVA uses this SV as an additional known factor in Step 1 of the next iteration. The algorithm stops, when no more significant hidden factor are detected in Step 3. Significant SVs obtained via IA-SVA can be used in subsequent analyses. For instance, in differential gene expression analyses SVs can be added as covariates in a regression model to adjust for the unwanted variation. If SVs explain biological variables of interest, e.g., cell type assignments, marker genes for SVs can be further utilized (e.g., marker genes for different cell types).

Assessing the significance of the contribution of a hidden factor in the variation of residuals. To assess the significance of a putative hidden factor (i.e., PC1 obtained from Step 2 in the previous section), we used the permutation based significance test applied in the surrogate variable analysis [5, 23]. Unlike SVA, which tests all putative hidden factors at once, IA-SVA assesses the significance of hidden factors one at a time during the corresponding iteration. Briefly, IA-SVA i) conducts SVD on the residual matrix obtained from Step 1, ii) computes the proportion of variation in this matrix explained by the first singular vector and iii) compares it against the values obtained from permuted residual matrices. The detailed steps of the algorithm are as follows:

[Step 1] Conduct SVD on the residual matrix.

[Step 2] Calculate the proportion of the variance in the residual matrix explained by the first singular vector using the test statistic: $T_{obs} = \frac{\lambda_1^2}{\sum_k \lambda_k^2}$, where λ_k is the *k*-th singular value.

[Step 3] Generate a permuted residual matrix by i) permuting each row of the log-transformed read count matrix *Y* and regressing *Y* on all known factors (*X* and *Z*) to obtain fitted residuals.

[Step 4] Repeat Step 3 *M* times and generate an empirical null distribution of the test statistics by calculating ($T_i^0, i = 1, \ldots, M$) for the *M* permuted residual matrices.

[Step 5] Compute the empirical p-value for the first singular vector (i.e., putative hidden factor) by counting the number of times the null statistics (T_i^0) exceeds the observed one (T_{obs}) divided by the number of permutations (*M*).

Gene expression data filtering. We filtered out low-expressed genes with read counts <= 5 in less than three cells and log-transformed the retained gene expression counts for further analyses.

Single-cell RNA-Seq data simulations. We simulated single-cell gene expression data with attributes similar to real-world scRNA-Seq data generated from human pancreatic islets [15]. We first estimated zero-inflated negative binomial model parameters (i.e., p_0: probabilities that the count will be zero, mu: mean of the negative binomial, size: size of the negative binomial) from this data using the Polyester R package [25]. With these model parameters, we simulated expression data for *m* expressed genes and *n* cells under two hypotheses: 1) the null hypothesis:

no hidden sources of variation, and 2) the alternative hypothesis: three hidden factors simulated in the data. Under both scenarios, we simulated a primary variable of interest (i.e., case vs. control) and simulated 10% of genes to be differentially expressed between the two groups. Under the alternative hypothesis, we simulated three hidden factors that affect 30%, 20% and 10% of randomly chosen genes respectively and simulated two different scenarios where these factors are moderately correlated (|*r*|=~0.3-0.6) or weakly correlated (|*r*|<0.3) with the group variable.

Detection power, Type I error rate and accuracy assessment. To assess the detection power, Type I error rate, and the accuracy of IA-SVA estimates, we simulated 1,000 times scRNA-Seq data (as explained in the previous section) for 10,000 genes and 50 cells, under the null hypothesis (i.e., a group (case/control) variable affecting 10% of genes and no hidden factor) and under the alternative hypothesis (i.e., a group variable and three hidden factors affecting 10%, 30%, 20%, 10% of genes, respectively). Under the alternative hypothesis, we considered two correlation scenarios where the three hidden factors are moderately (|*r*|=~0.3-0.6) or weakly (|*r*|<0.3) correlated with the group variable. We used 0.05 as the nominal significance level (α). Accordingly, for USVA and SSVA analyses, we set α at 0.05 by modifying the *'num.sv'* function in the svaseq R package[4]. 50 permutations were used to test the significance of a factor's contribution to the unexplained variation in the data. We defined the empirical Type I error rate as the number of times each method detects a false positive factor under the null hypothesis (i.e., a factor does not exist but is detected as significant at the nominal p-value threshold of 0.05) divided by the number of simulations (i.e., 1,000). Similarly, the empirical power rate for detecting a hidden factor is defined as the number of times each method detects a simulated factor under the alternative hypothesis (i.e., a factor actually exists and is detected as significant by the method) divided by 1,000. We assessed the accuracy of the estimates using the average of the absolute correlation coefficients between the simulated and estimated hidden factors.

Inference of cell types from brain cells. For a more realistic assessment of algorithms, we used gene expression profiles of neurons (n=52) and oligodendrocytes (n=20) obtained from two different brain tissues: cortex (n=65) and hippocampus (n=7) [14]. We treated the cell type assignments (neuron vs. oligodendrocyte) as an unknown variable and estimated it by computing the top SV (or PC in case of PCA) using IA-SVA and other unsupervised methods (i.e., USVA, PCA, RUVemp and RUVres). Given that neurons and oligodendrocytes have very different expression profiles, if entire genes are used for this analysis, all methods will deliver perfect estimates. Thus, to enable performance comparisons, we made the problem more challenging by randomly choosing 1,000 genes and considering only these genes in the analyses (same random set of genes used for all methods for comparability). The number of expressed genes in each cell is a major source of cell-to-cell variation in scRNA-Seq data and frequently correlates with other factors [12]. Thus, 'Sample ID' and the number of expressed genes are included into IA-SVA, USVA and RUVres models as known factors. We assessed the accuracy of each method in inferring the true cell type by calculating the absolute Pearson correlation coefficient ($|r|$) between inferred cell types and an indicator variable for the true cell type (e.g., taking one for neurons and zero for oligodendrocytes).

Detection of a subset of alpha cells that uniquely express a subset of genes. To test whether IA-SVA is effective in capturing heterogeneity within a relatively homogenous cell population, we studied islet alpha cells (n=101) from three diabetic patients [15]. After filtering weakly expressed genes, 14,416 genes out of 26,616 were used for further analyses. 'Patient ID' and geometric library size are modeled as known factors, and top 3 significant factors contributing to the unexplained variation are inferred using IA-SVA at p-value of 0.05 using 50 permutations. For comparison, we applied PCA, USVA, and tSNE on this data. In the USVA analysis, we similarly used 'Patient ID' and the geometric library size as known factors. In the PCA analysis, PC1 is discarded since it is highly correlated with the number of expressed genes. To test whether the heterogeneity detected in alpha cells is reproducible, we conducted similar analyses on a bigger human islet scRNA-Seq dataset independently generated with the Fluidigm C1 platform [16]. We used gene expression profiles of 563 alpha cells from six diabetic patients. After removing weakly expressed genes, 17,025 genes were retained. 'Patient ID' and the geometric library size are modeled as known factors in our models, and top 3 significant SVs are obtained using IA-SVA. For comparison, we conducted similar analyses using PCA (PC1 and PC2 are discarded since PC1 matched number of expressed genes and PC2 captured the 'Patient ID', which are adjusted for in IA-SVA and USVA), USVA and tSNE. For USVA, similarly, we adjusted for 'Patient ID' and the geometric library size.

Detection of heterogeneity stemming from cell-cycle stage differences. To assess the performance of IA-SVA and existing methods in detecting the effect of cell-cycle stage, we analyzed scRNA-Seq data obtained from human glioblastomas, which has an established cell-cycle signature [19]. We considered gene expression read counts of 25,415 genes and 58 cells obtained from a tumor sample (MGH30). After filtering out lowly expressed genes, 21,151 genes were retained. Using IA-SVA, we adjusted for geometric library size at the initial step and iteratively extracted top 3 significant SVs at p-value of 0.05 using 50 permutations. For comparison, we applied PCA, USVA and tSNE on this data. In USVA, similarly, we adjusted for geometric library size.

IA-SVA based gene selection can improve the performance of clustering algorithms. To compare the performance of tSNE combined with IA-SVA against standard tSNE analyses, we studied gene expression profiles of alpha (n=101, marked with glucagon (*GCG*) expression), beta (n=96, marked with insulin (*INS*) expression), and ductal (n=16, marked with *KRT19* expression) cells obtained from three diabetic patients [15]. We filtered out low-expressed genes and retained 16,047 genes for further analyses. Then, we performed IA-SVA based marker gene selection and conducted tSNE on these selected genes. For comparison we also performed tSNE on all expressed genes (n=16,047). We repeated similar analyses on a bigger and more complex data generated using Fluidigm C1 platform [16], which contains 1,600 cells (alpha (n=946), beta (n=503), delta (n=58) and PP (n=93)) obtained from 6 diabetic and 12 non-diabetic individuals. After filtering lowly expressed genes, the number of retained genes was 19,226. We first clustered these 1,600 cells by performing tSNE on all expressed genes (n=19,226). Next, we conducted IA-SVA analyses while accounting for the known factors (i.e., Patient ID, Phenotype (diabetic vs. non-diabetic), sex and geometric library size) and performed tSNE analysis on the marker genes inferred by IA-SVA.

References

1. Tung, P.-Y. et al. Batch effects and the effective design of single-cell gene expression studies. *bioRxiv*, 062919 (2016).
2. Kowalczyk, M.S. et al. Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome research* **25**, 1860-1872 (2015).
3. Stegle, O., Teichmann, S.A. & Marioni, J.C. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* **16**, 133-145 (2015).
4. Leek, J.T. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res* **42** (2014).
5. Leek, J.T. & Storey, J.D. A general framework for multiple testing dependence. *Proc Natl Acad Sci U S A* **105**, 18718-18723 (2008).
6. Teschendorff, A.E., Zhuang, J. & Widschwendter, M. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics* **27**, 1496-1505 (2011).
7. Risso, D., Ngai, J., Speed, T.P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol* **32**, 896-902 (2014).
8. Gagnon-Bartsch, J.A. & Speed, T.P. Using control genes to correct for unwanted variation in microarray data. *Biostatistics* **13**, 539-552 (2012).
9. Buettner, F. et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotechnol* **33**, 155-160 (2015).
10. Ilicic, T. et al. Classification of low quality cells from single-cell RNA-seq data. *Genome Biol* **17**, 29 (2016).
11. McDavid, A., Finak, G. & Gottardo, R. The contribution of cell cycle to heterogeneity in single-cell RNA-seq data. *Nat Biotechnol* **34**, 591-593 (2016).
12. Hicks, S.C., Teng, M. & Irizarry, R.A. On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data. *bioRxiv* (2015).
13. Maaten, L.V.D. Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.* **15**, 3221-3245 (2014).
14. Darmanis, S. et al. A survey of human brain transcriptome diversity at the single cell level. *Proc Natl Acad Sci U S A* **112**, 7285-7290 (2015).
15. Lawlor, N. et al. Single cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res* (2016).
16. Xin, Y. et al. RNA Sequencing of Single Human Islet Cells Reveals Type 2 Diabetes Genes. *Cell Metab* **24**, 608-615 (2016).
17. Xin, Y. et al. Use of the Fluidigm C1 platform for RNA sequencing of single mouse pancreatic islet cells. *Proc Natl Acad Sci U S A* **113**, 3293-3298 (2016).
18. Wang, Y.J. et al. Single-Cell Transcriptomics of the Human Endocrine Pancreas. *Diabetes* **65**, 3028-3038 (2016).
19. Patel, A.P. et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396-1401 (2014).
20. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* **44**, D457-462 (2016).
21. Gene Ontology, C. Gene Ontology Consortium: going forward. *Nucleic Acids Res* **43**, D1049-1056 (2015).

22. Macosko, E.Z. et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).

23. Buja, A. & Eyuboglu, N. Remarks on Parallel Analysis. *Multivariate Behav Res* **27**, 509-540 (1992).

24. Leek, J.T. & Storey, J.D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* **3**, 1724-1735 (2007).

25. Frazee, A.C., Jaffe, A.E., Langmead, B. & Leek, J.T. Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics* **31**, 2778-2784 (2015).