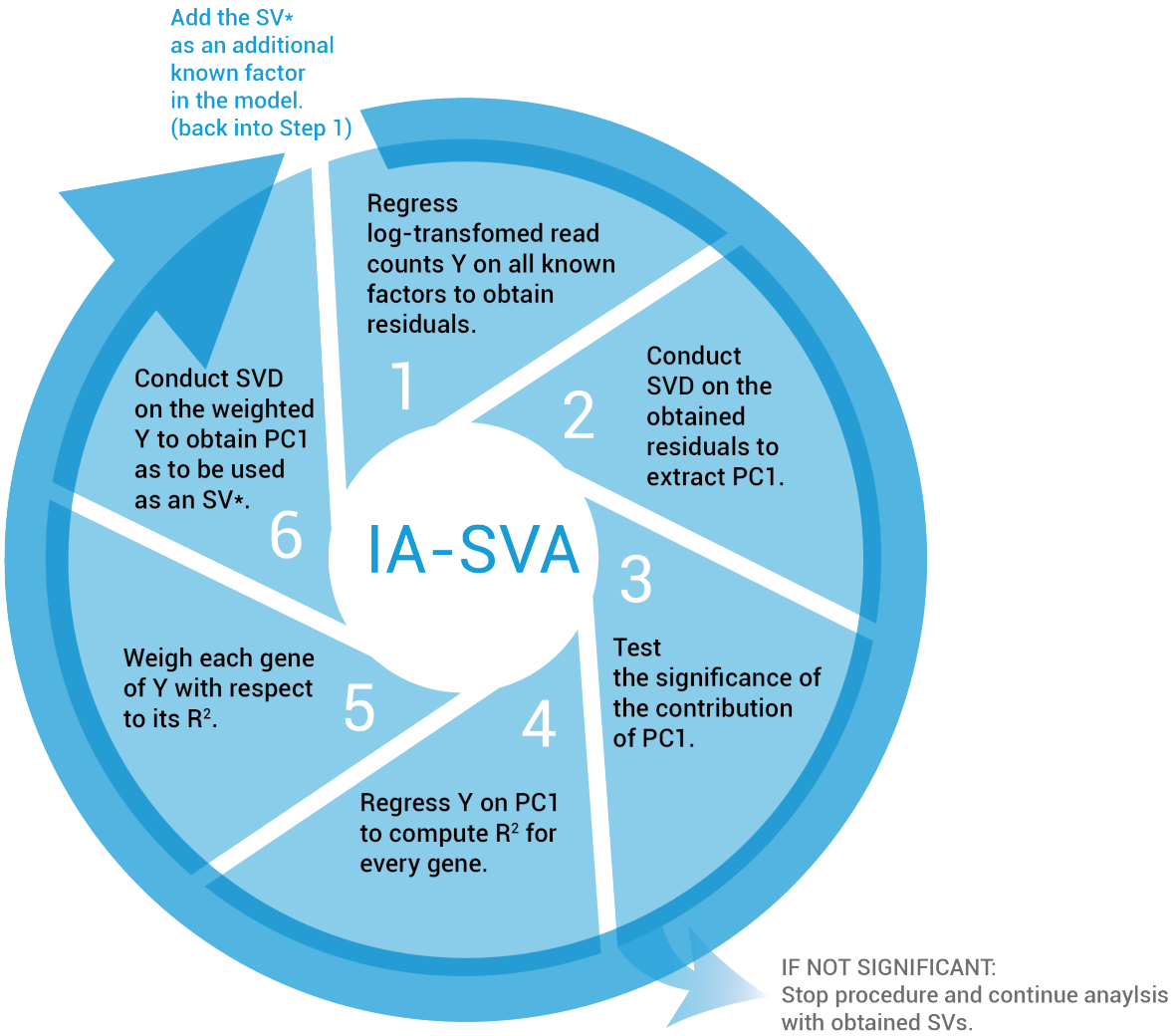
A robust statistical framework to detect multiple sources of hidden variation in single-cell transcriptomes

Supplementary Information

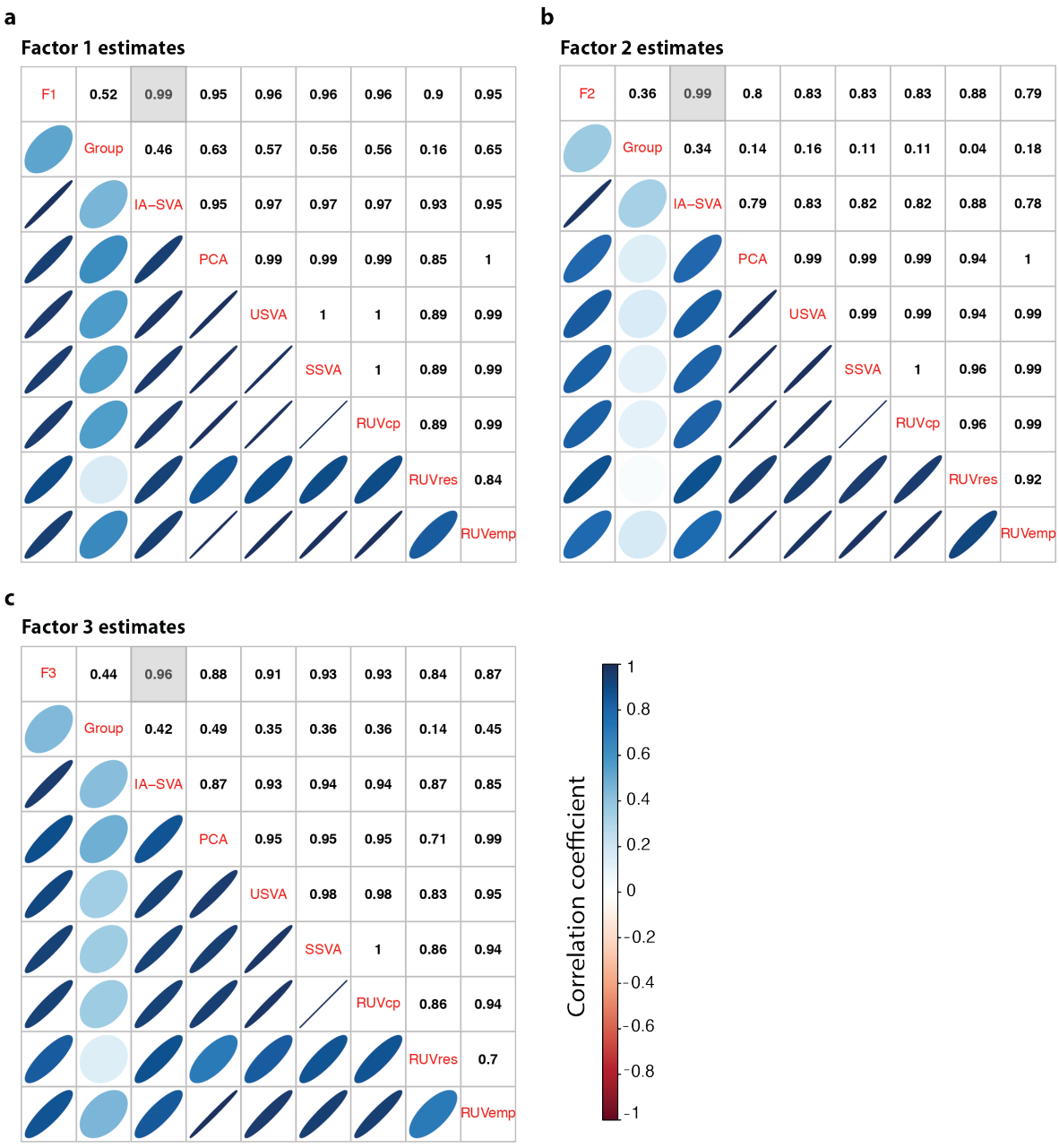
Donghyung Lee[1,*], Anthony Cheng[1,2] and Duygu Ucar[1,*]

[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

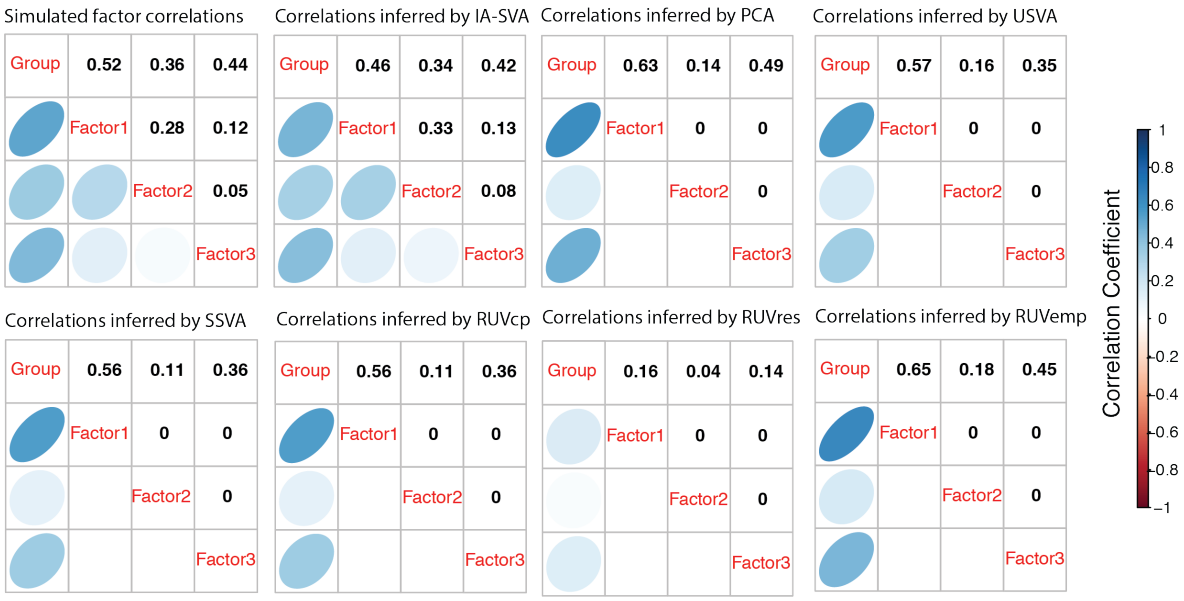
*To whom correspondence should be addressed.



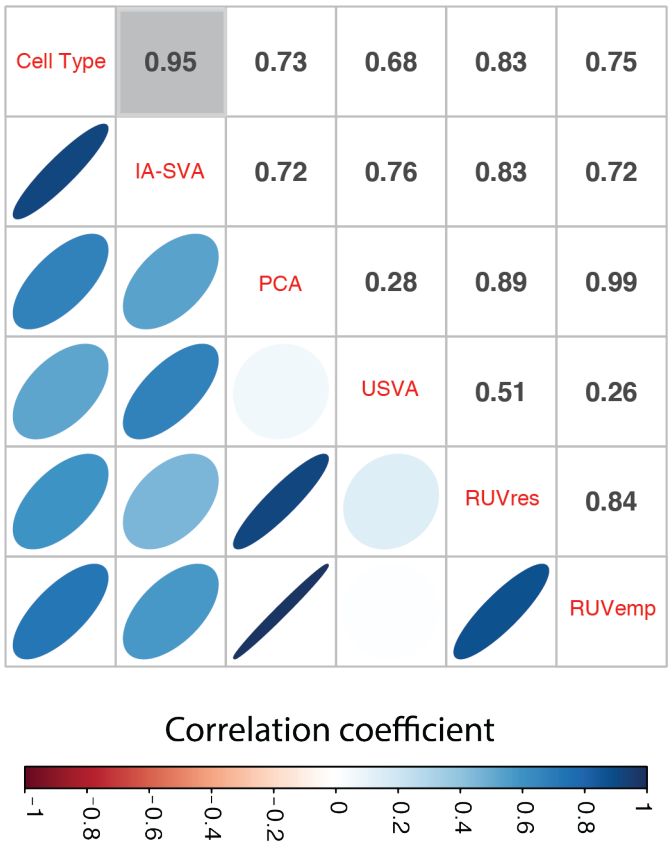
Supplementary Figure 1. **Summary of the IA-SVA framework.** At each iteration IA-SVA computes the first principal component (PC1) from read counts adjusted for all known factors and tests its significance [Steps 1-3]. If significant, IA-SVA uses this PC1 to infer marker genes associated with the hidden factor [Steps 4-5] and obtain a surrogate variable (SV) to represent the hidden factor using these marker genes [Step 6]. In the next iteration, IA-SVA uses the obtained SV as an additional known variable to find further hidden factors.



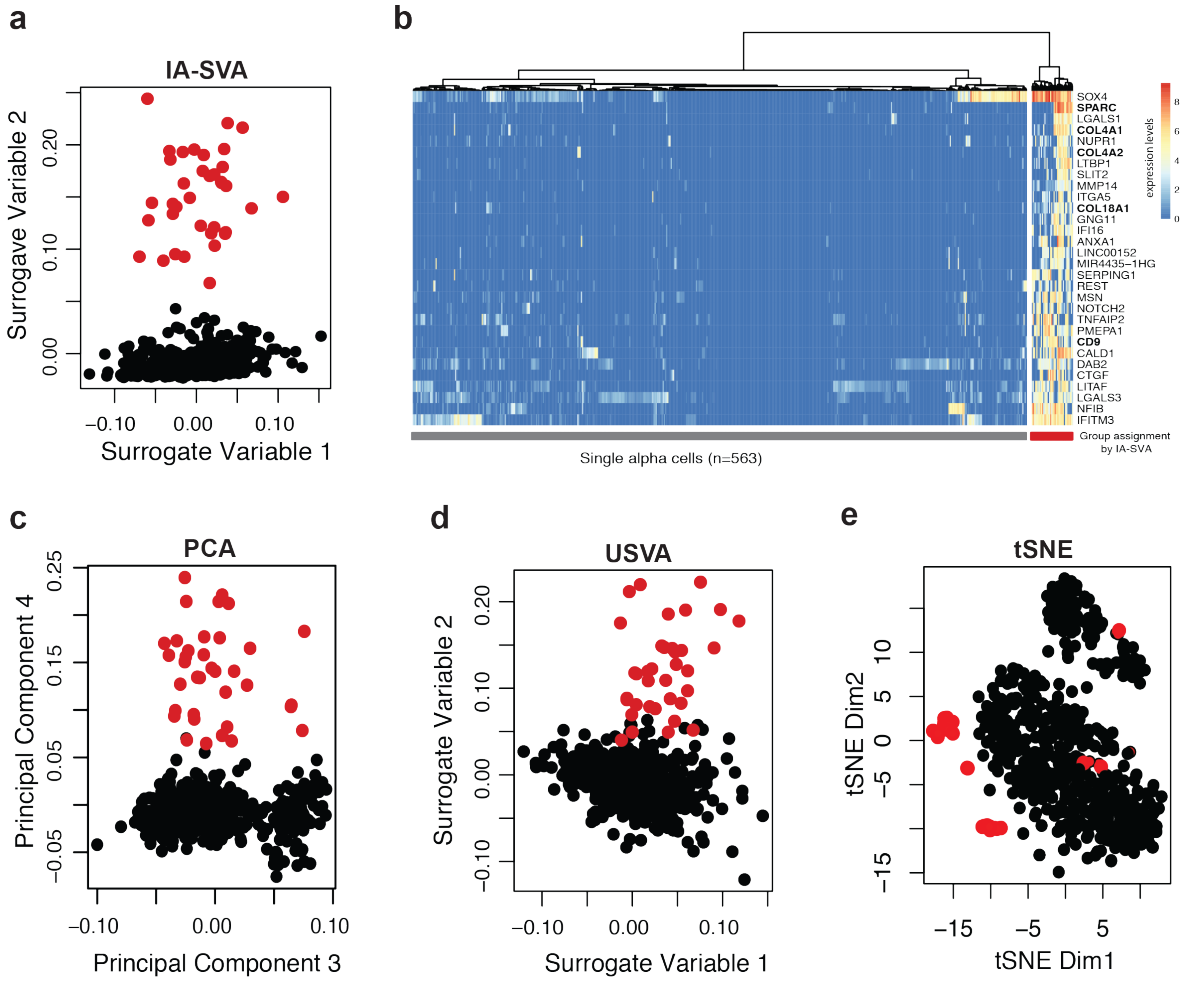
Supplementary Figure 2. **Pearson correlation coefficients between simulated factor, group variable, and factor estimates from IA-SVA and alternative supervised (SSVA and RUVcp) and unsupervised (USVA, PCA, RUVemp and RUVres) methods.** IA-SVA outperforms existing methods in terms of accuracy of the estimate for all three factors. See **Supplementary Note 1** for more details of the data simulation and experiments.



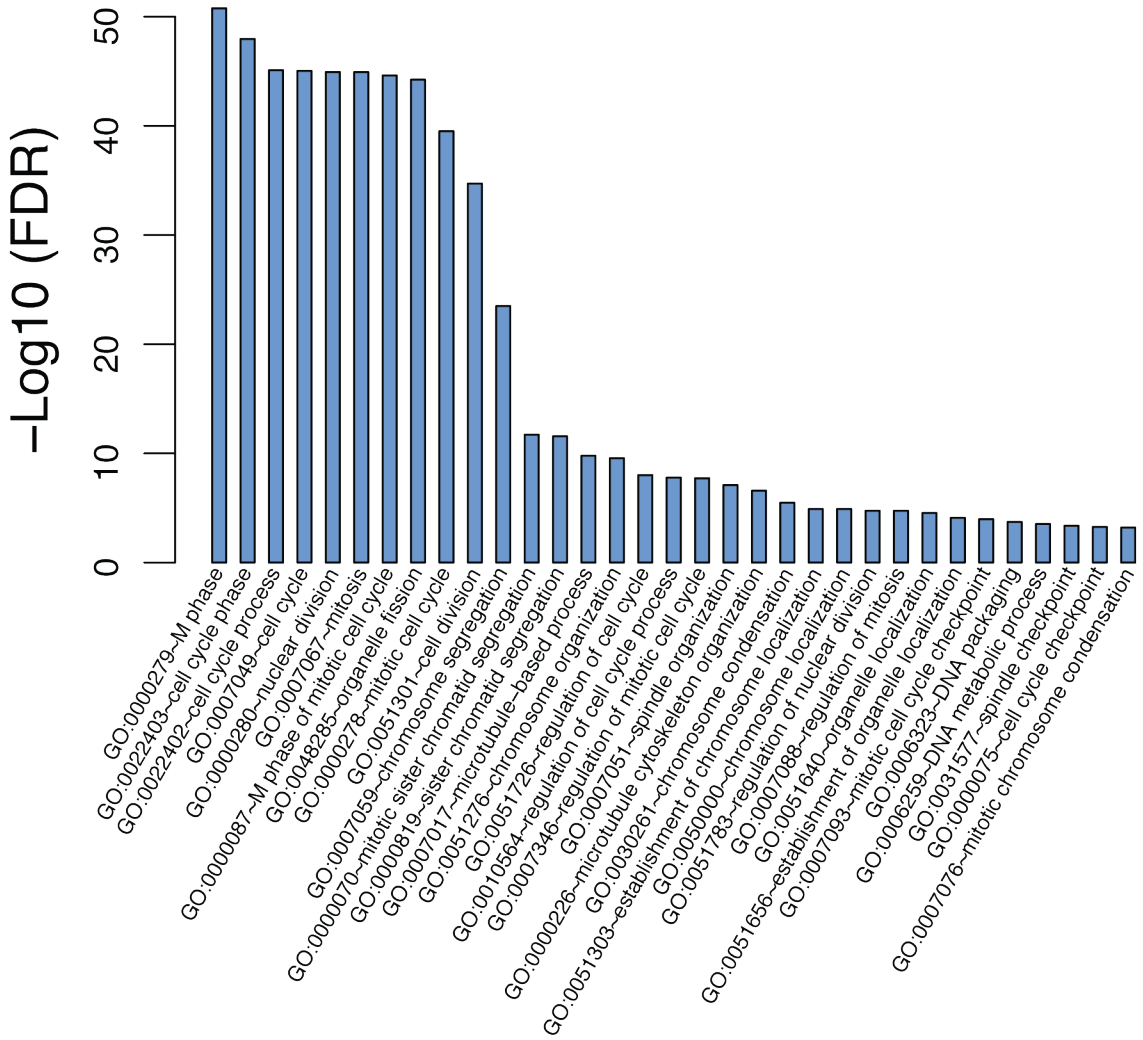
Supplementary Figure 3. Correlation structure among true and estimated factors (Group, Factor1, Factor2 and Factor3) and the group variable based on simulated scRNA-seq data. We studied the true correlation structure (Pearson correlation coefficient) among all simulated factors (Group, Factor1, Factor2 and Factor3) and compared this against the correlation structure based on detected factors. IA-SVA accurately estimated correlations between the group variable and hidden factors, whereas other methods failed to do so particularly for the correlations between three hidden factors due to their orthogonality assumption.



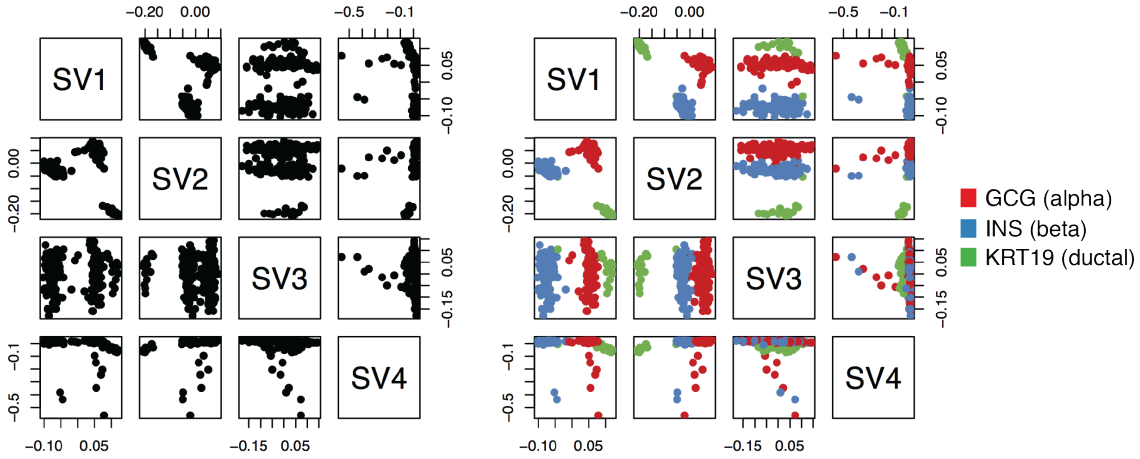
Supplementary Figure 4. The absolute correlation coefficient between the cell type estimates obtained via different methods and the indicator variable of the true cell type assignments in brain scRNA-Seq data. IA-SVA outperforms alternatives in inferring cell type as a hidden factor. Upper half of the matrices represent the absolute correlation coefficient scores for each comparison, where as lower half depicts the strength of the absolute correlation scores. The correlation score for IA-SVA's estimate is marked with gray box.



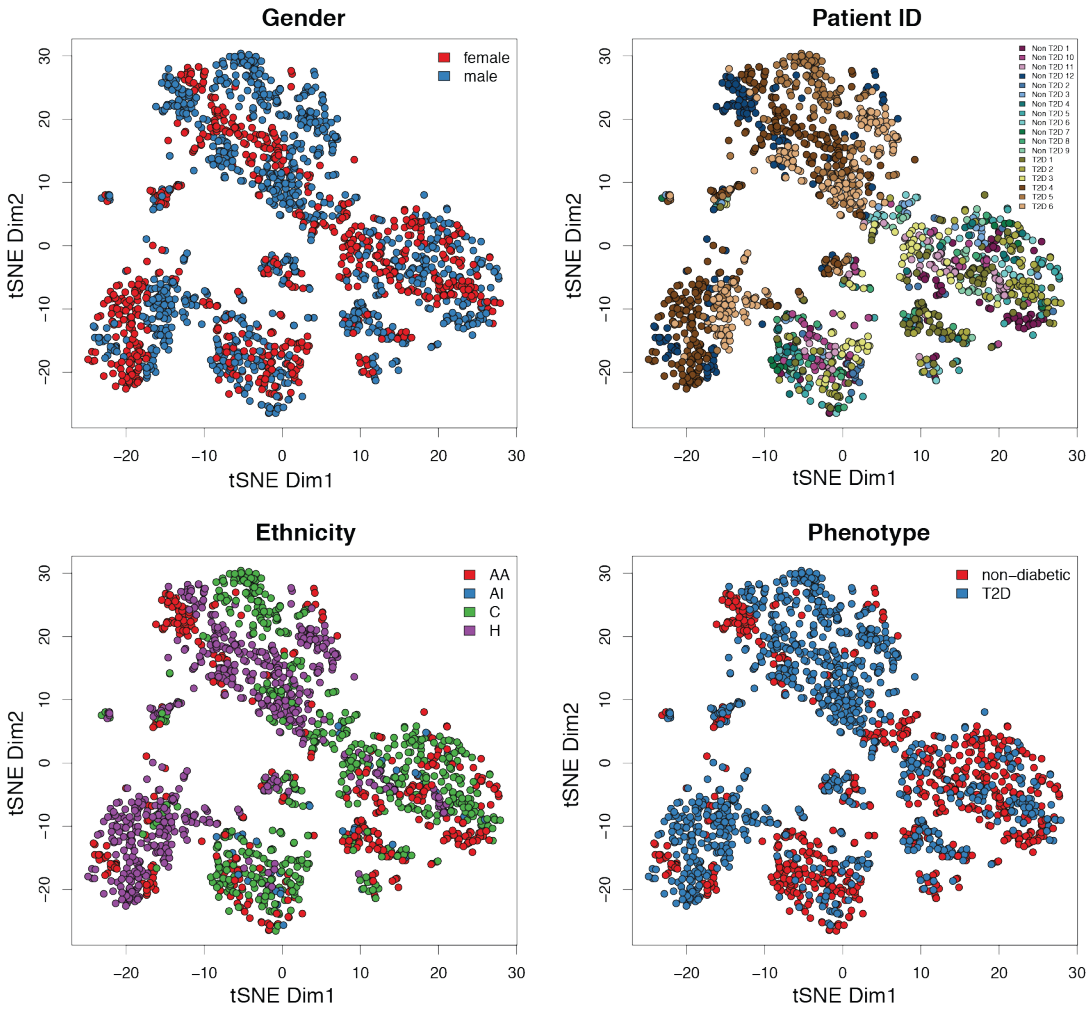
Supplementary Figure 5. IA-SVA recapitulates detected heterogeneity in alpha cells in a second pancreatic islet scRNA-seq data. **(a)** Heterogeneity captured within alpha cells using IA-SVA. Cells are clustered into two groups (black vs. red dots) based on IA-SVA's surrogate variable 2 (SV2 > 0.05). **(b)** Hierarchical clustering of alpha cells using the top 30 marker genes (ward.D2 and cutree_cols =2). 36 cells clearly separate from the rest of the cells based on their high expression of these genes. Heterogeneity captured by (**c**) PCA, (**d**) USVA and (**e**) tSNE. In PCA, PC1 and PC2 were disregarded since PC1 and PC2 map to the number of expressed genes and patient id respectively. While PCA, USVA and tSNE detected some heterogeneity among alpha cells, they failed to clearly separate these 36 cells. PCA and tSNE captured clusters originated from known factors (e.g., patient id), which are adjusted for in IA-SVA and USVA.



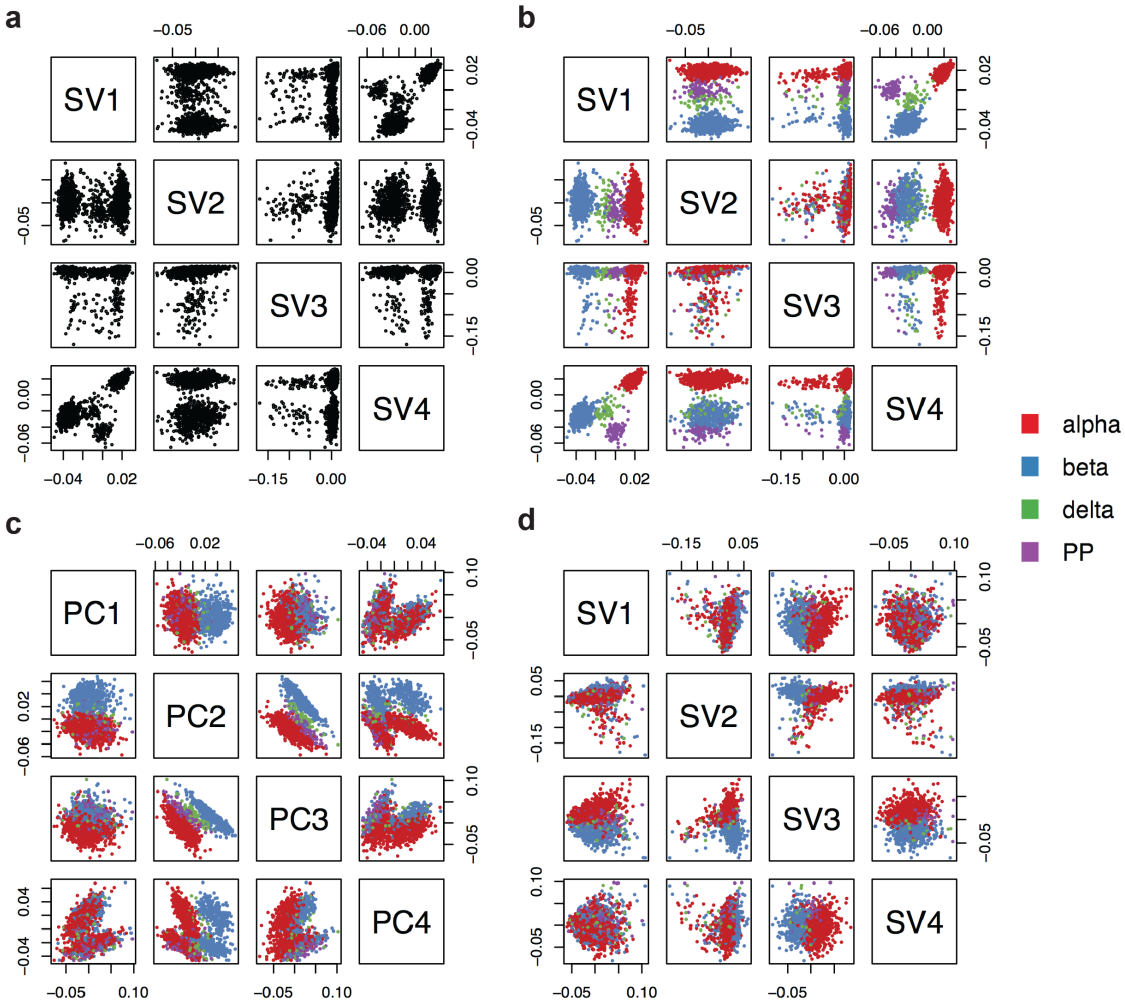
Supplementary Figure 6. GO term enrichment analyses for 87 marker genes associated with IA-SVA's SV2. FDR corrected enrichment p-values are depicted. Note that these genes are strongly enriched in cell-cycle related GO terms, where 86 out of 87 genes are associated with at least one such GO term.



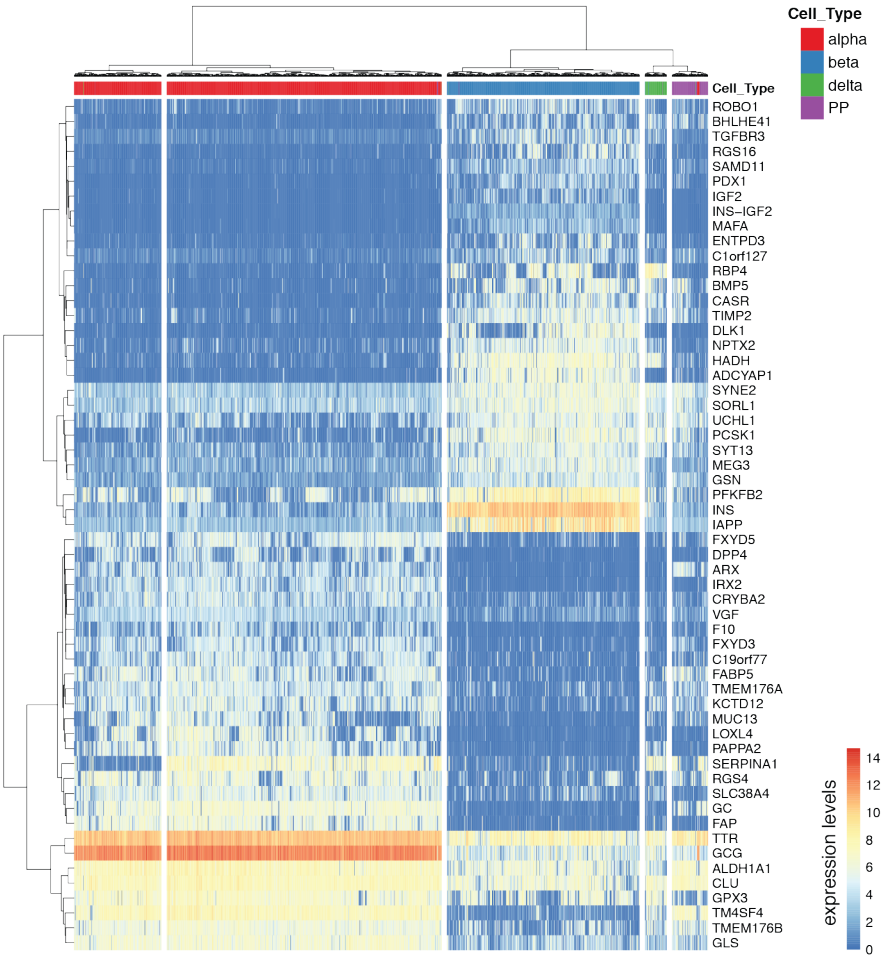
Supplementary Figure 7. **Pairwise scatter plot of top four significant IA-SVA surrogate variables (SV) detected from human islet scRNA-Seq data including three cell types: alpha (GCG), beta (INS) and ductal (KRT19) cells.** Cells on the right subfigure are color-coded based on the original assignment. SV1 and SV2 clearly separate cells into distinct clusters, therefore are good candidates for further analyses. SV4 captures technical heterogeneity stemming from cell contamination (e.g., stacked doublets), which was observed in **Figure 2** and **Supplementary Figure 5**.



Supplementary Figure 8. **Known variables explain single cell clustering and may confound with the heterogeneity stemming from different cell types.** tSNE plots generated using entire set of expressed genes are color-coded using known variables: sex, patient id, ethnicity, and phenotype. Among these, patient id and ethnicity drive the clustering of cells and can lead to misinterpretations of cell types.



Supplementary Figure 9. **IA-SVA effectively dissects the hidden variation in a second human islet scRNA-Seq data with strong confounders. (a)** Pairwise scatter plot of top four significant IA-SVA surrogate variables (SV). **(b)** Same as panel (a) where cell are color-coded with respect to original cell assignments. SV1 and SV4 separate cells into disjoint clusters that matches to respective cell types as determined in the original study. SV3 captures technical heterogeneity stemming from stacked doublet cells, which was observed in **Figure 2** and **Supplementary Figure 5**. **(c)** Pairwise scatter plot of top four PCs from PCA on the same data. **(d)** Pairwise scatter plot of top four significant SVs obtained from USVA adjusted for all known factors that are also considered in the IA-SVA analysis (i.e., patient id, phenotype, sex and geometric library size). IA-SVA outperforms alternatives in capturing hidden factors associated with cell types.



Supplementary Figure 10. IA-SVA detects marker genes associated with different cell types among islet cells. Hierarchical clustering of islet cells using 57 marker genes detected by IA-SVA (ward.D2 and cutree_cols = 4). These genes are significantly associated (Benjamini-Hochberg q-value < 0.05) and highly correlated ($r^2 > 0.3$) with IA-SVA's SV1 and SV4. Note that cells are clustered together based on their cell types. Color-coding is based on the original study's assignments.

Supplementary Note 1. IA-SVA outperforms existing methods for detecting hidden sources of variation.

We simulated gene expression levels for 2,000 genes and 50 cells under the alternative hypothesis (i.e., existence of hidden factors) with the moderate correlation scenario ($|r|$=0.3~0.6). We applied IA-SVA and supervised (SSVA [1] and RUVcp[2]) and unsupervised (USVA[1], PCA, RUVemp [2] and RUVres [2]) methods to detect the simulated factors . Among the studied methods, only IA-SVA, USVA and SSVA can infer the number of hidden factors in the data (k). For other methods, we used $k=3$. For the significance assessment of estimated factors we used 20 permutations and set the nominal level of significance at 0.05 for IA-SVA, USVA and SSVA analyses. While all three (IA-SVA, USVA and SSVA) determined the number of hidden factors correctly (i.e., k=3), IA-SVA outperformed other methods in terms of the accuracy of the estimate (the correlation between simulated and estimated factors) (**Supplementary Fig. 2**). For the estimation of Factor1 (i.e., the factor that affects 30% of genes), IA-SVA outperformed all tested supervised and unsupervised methods where the correlation between the IA-SVA estimation and the simulated factor was 0.99 (**Supplementary Fig. 2a**). The efficacy of IA-SVA was more evident for factors affecting smaller number of genes. For example, for Factor 2 (affecting 20% of genes), the IA-SVA estimate was almost perfectly correlated with the true factor (r=0.99), whereas the correlation was lower for the other two methods (r=0.83) (**Supplementary Fig. 2b**).

References

1. Leek, J.T. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res* **42** (2014).
2. 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).