A statistical framework for the robust detection of hidden variation in single-cell transcriptomes

Supplementary Information

Donghyung Lee[1,*], Anthony Cheng[1,2], Mohan Bolisetty[1], and Duygu Ucar[1,3,*]

[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, [3] Institute of Systems Genomics, University of Connecticut Health Center, Farmington, CT, USA.

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

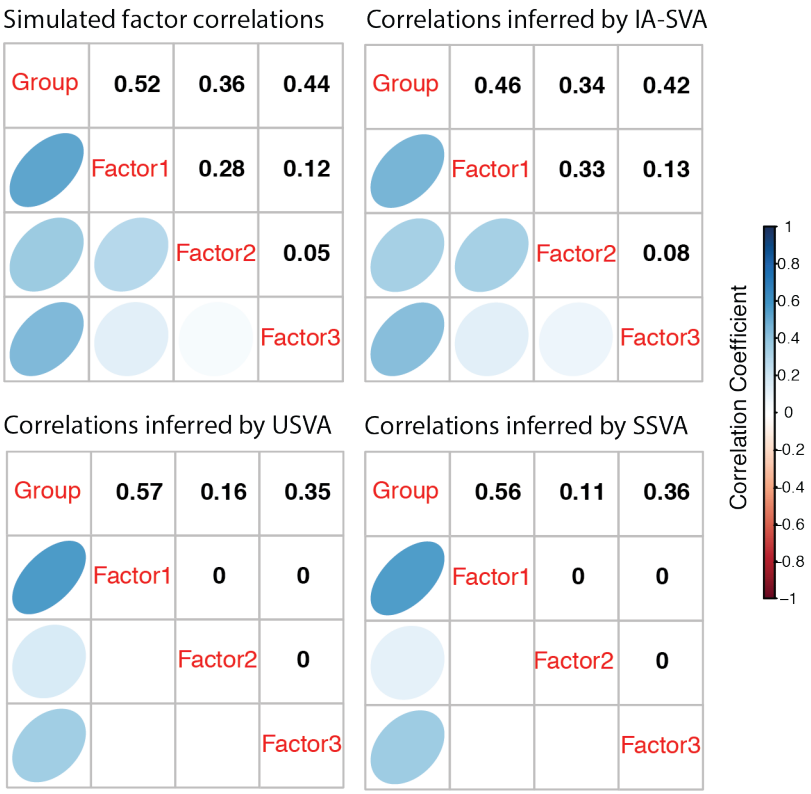

Figure S1. **Correlation structure among true and estimated factors (Group, Factor1, Factor2 and Factor3) and the group variable based on simulated scRNA-seq data.** We studied the true correlation structure (Pearson correlation coefficient) among all simulated factors (Group, Factor1, Factor2 and Factor3) and compared this against the correlation structure based on detected factors. IA-SVA accurately estimated correlations between the group variable and hidden factors, whereas SVA methods failed to do so particularly for the correlations between three hidden factors due to their orthogonality assumption.

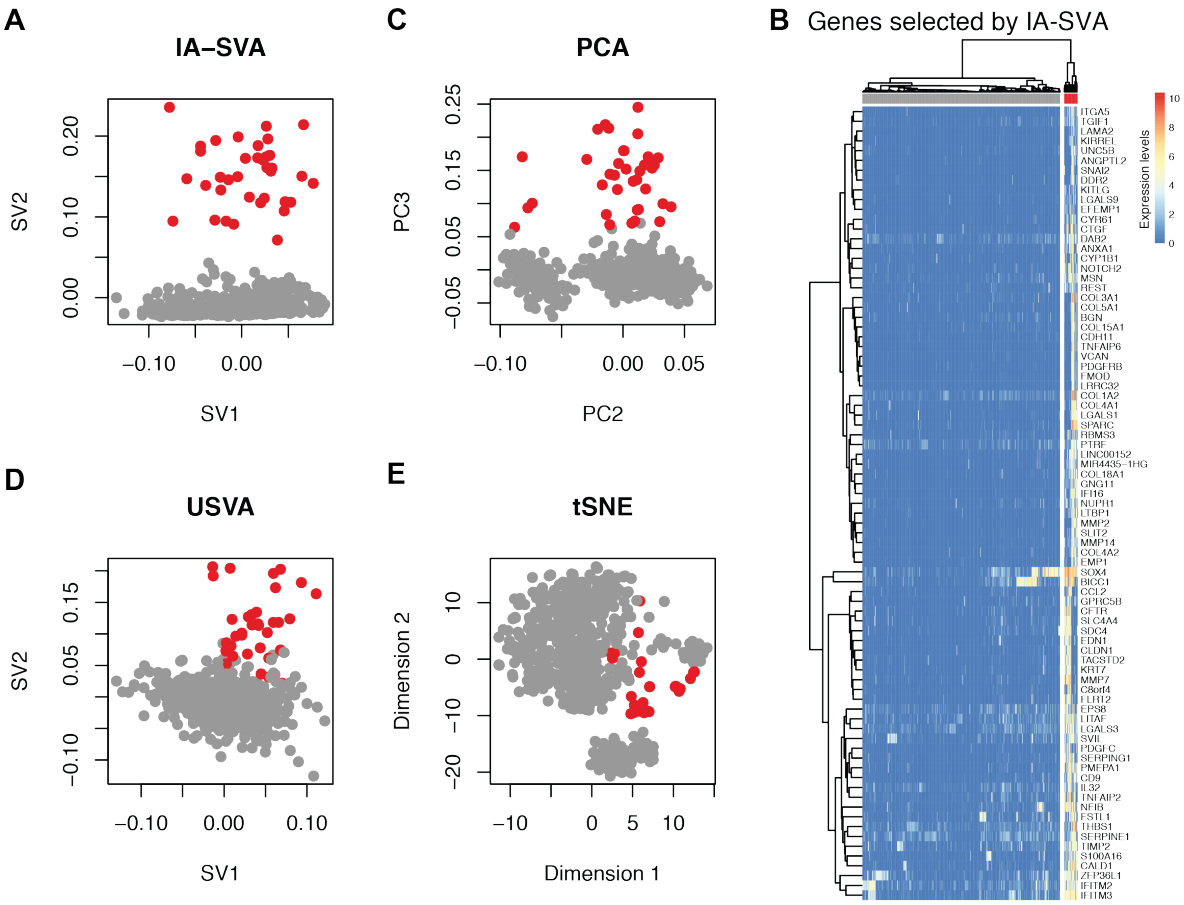

Figure S2 IA-SVA recapitulates detected heterogeneity in alpha cells in a second pancreatic islet scRNA-seq data. (A) Outlier alpha cells captured using IA-SVA and same cells marked in respective (**C**) PCA, (**D**) USVA, and (**E**) tSNE analyses. Cells are clustered into two groups (red vs. gray dots) based on IA-SVA's surrogate variable 2 (SV2 > 0.05). **(B)** Hierarchical clustering (ward.D2 and cutree_cols =2) of alpha cells using 81 genes significantly associated (FDR < 0.05 and R^2 > 0.3) with SV2. 36 cells clearly separate from the rest of the cells based on their high expression of these genes. In PCA, PC1 was disregarded since it maps to the geometric library size. 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.

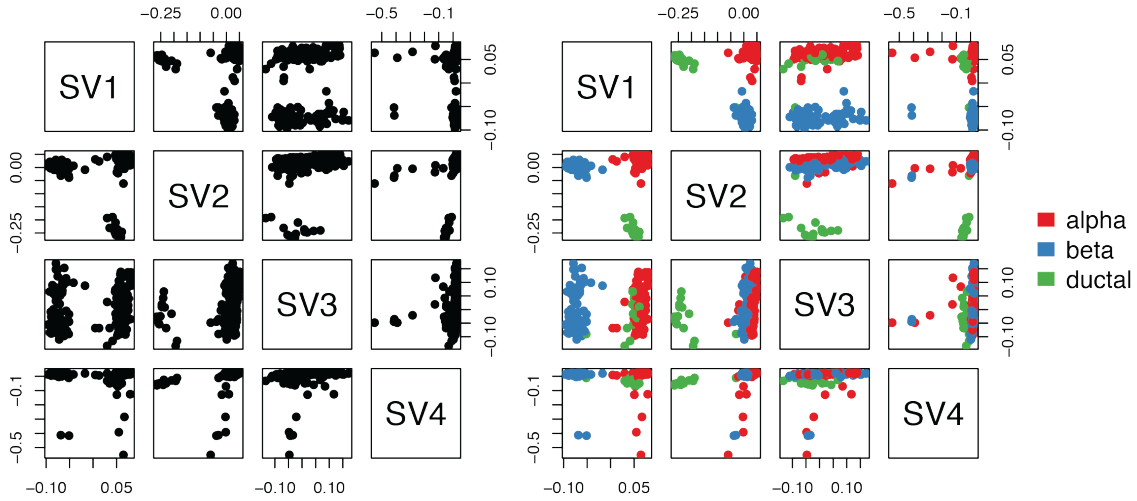

Figure S3. **Pairwise scatter plot of top four significant IA-SVA surrogate variables (SV) detected from human islet scRNA-Seq data including three cell types: alpha (GCG), beta (INS) and ductal (KRT19) cells.** Cells on the right subfigure are color-coded based on the original assignment. SV1 and SV2 clearly separate cells into distinct clusters, therefore are good candidates for further analyses. GO enrichment and pathway analyses results of 92 genes associated (FDR < 0.05, $R^2 > 0.5$) with SV1 and SV2 are illustrated in Supplementary Table S4. SV4 captures technical heterogeneity stemming from cell contamination (e.g., stacked doublets), which was observed in **Figure 2** and **Figure S2**. Go enrichment and pathway analyses results of 94 genes associated (FDR < 0.05, $R^2 > 0.3$) with SV4 are illustrated in Supplementary Table S6.

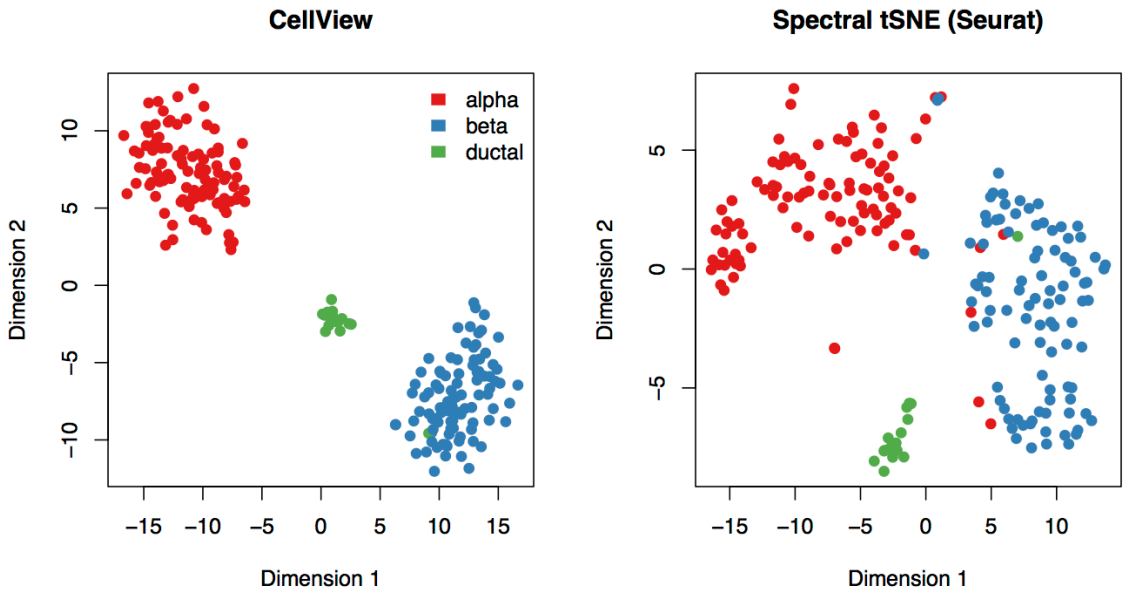

Figure S4. **tSNE analyses using 1000 most over-dispersed genes (CellView, left) and using significant PCs obtained from highly over-dispersed genes detected Seurat (Spectral tSNE, right)**. Cells are color-coded based on the original cell-type assignment.

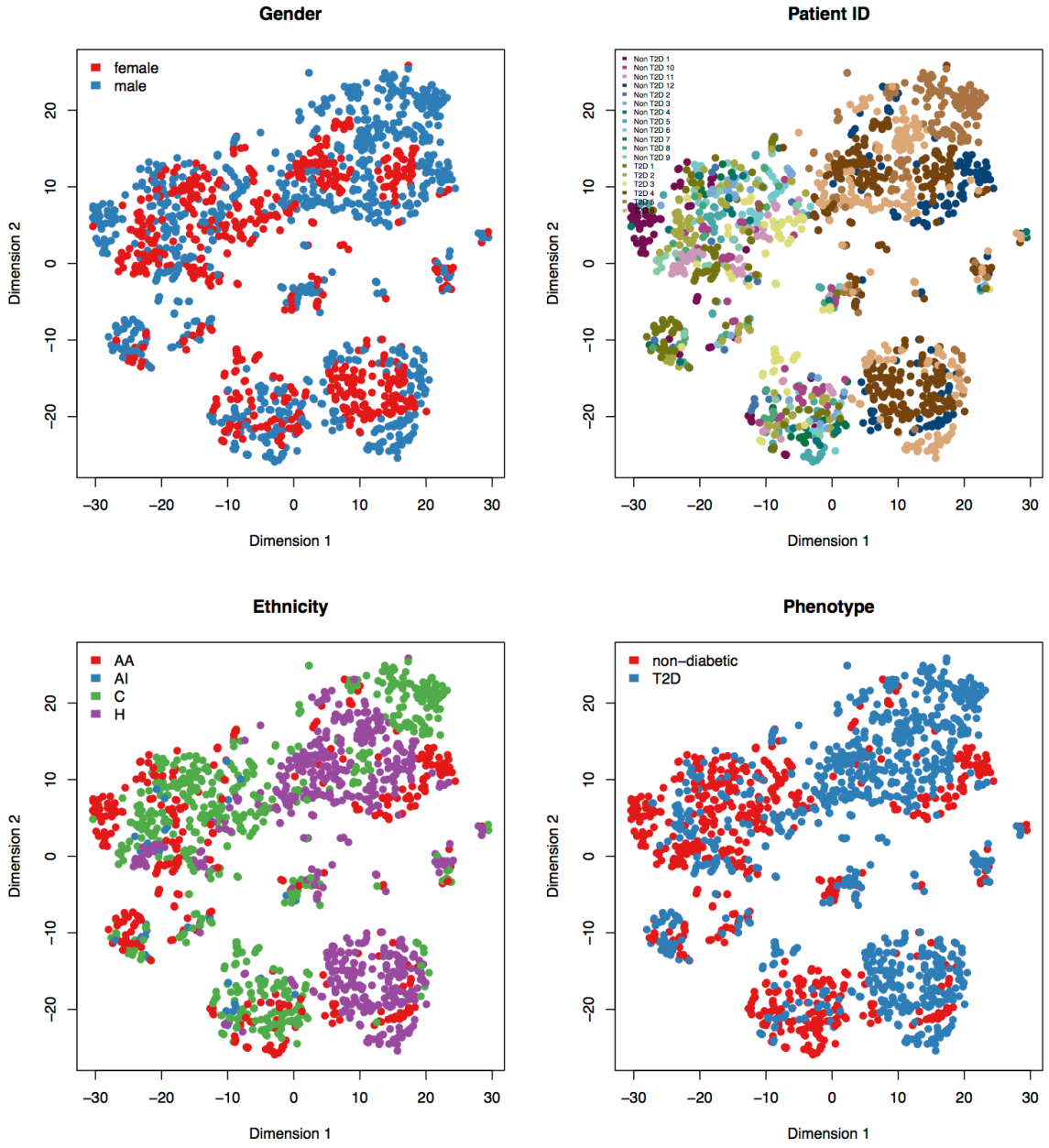

Figure S5. **Known variables explain single cell clustering and may confound with the heterogeneity stemming from different cell types.** tSNE plots generated using entire set of expressed genes are color-coded using known variables: sex, patient ID, ethnicity, and phenotype. Among these, patient ID and ethnicity drive the clustering of cells and can lead to misinterpretations of cell types.

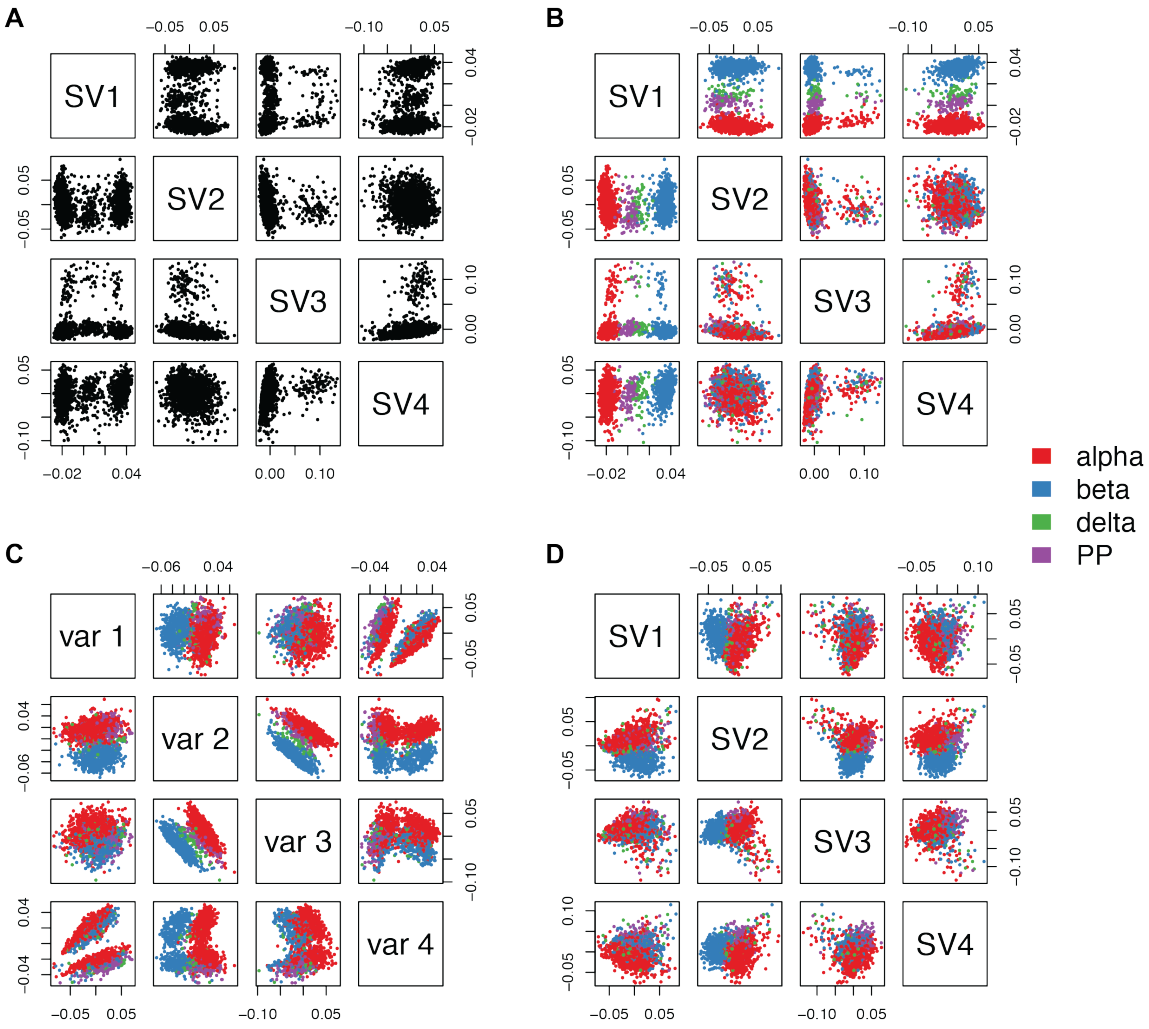

Figure S6. **IA-SVA effectively dissects the hidden variation in a second human islet scRNA-Seq data with strong confounders. (A)** Pairwise scatter plot of top four significant IA-SVA surrogate variables (SV). **(B)** Same as panel (**A**) where cell are color-coded with respect to original cell assignments. SV1 separate cells into disjoint clusters that matches to respective cell types as determined in the original study (see GO enrichment and pathway analyses results of 57 SV1 genes in Supplementary Table S5). SV3 captures technical heterogeneity stemming from stacked doublet cells (GO enrichment and pathway analyses results of 54 SV3 genes (FDR < 0.05 and $R^2 > 0.3$) are illustrated in Supplementary Table S7), which was observed in **Figure 2** and **Figure S2**. **(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 and geometric library size). IA-SVA outperforms alternatives in capturing hidden factors associated with cell types.

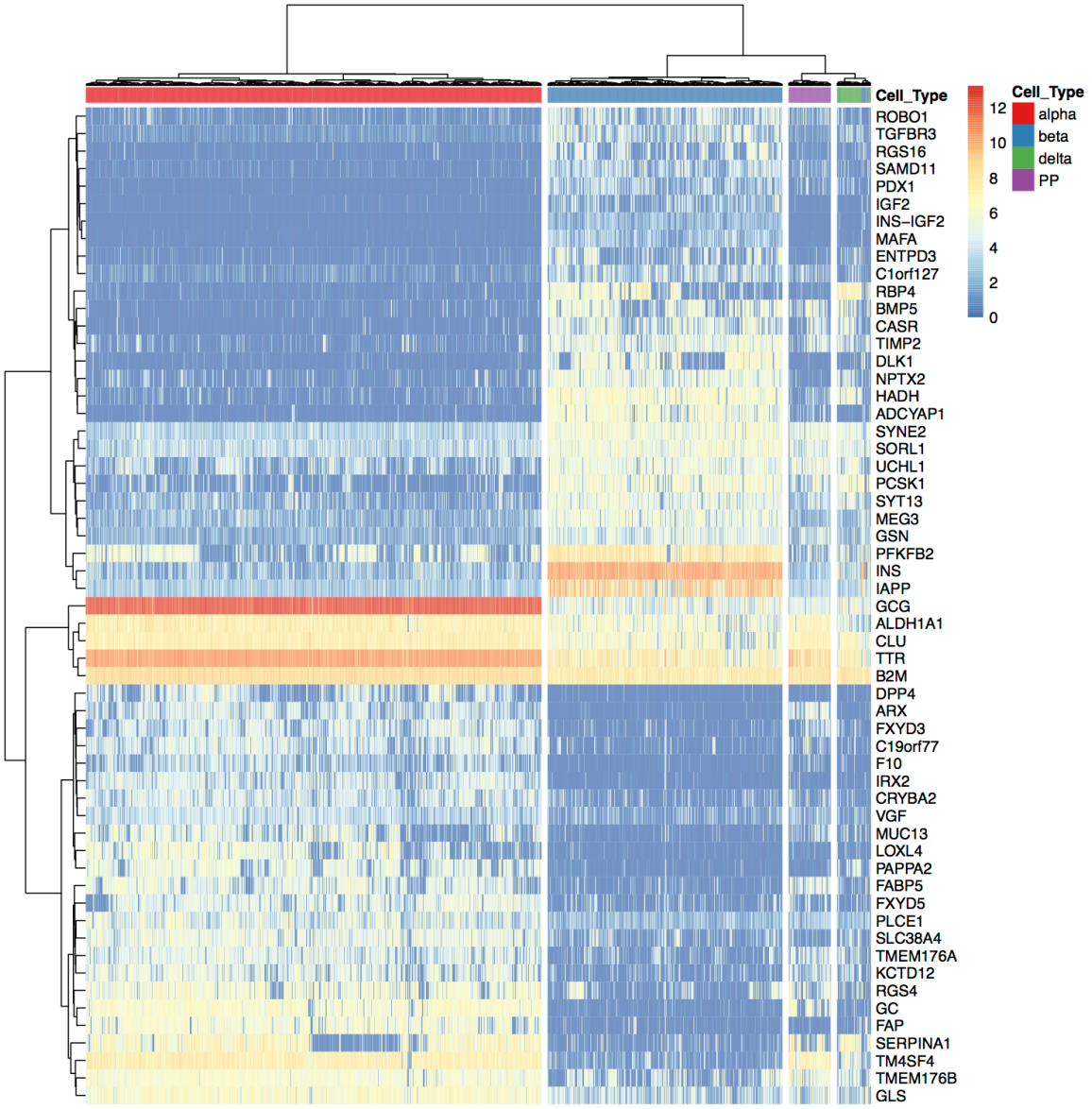

Figure S7. IA-SVA detects marker genes associated with different cell types among islet cells. Hierarchical clustering of islet cells using 57 marker genes detected by IA-SVA (ward.D2 and cutree_cols = 4). These genes are significantly associated (FDR < 0.05 and R^2 > 0.5) with IA-SVA's SV1. Note that cells are clustered together based on their cell types. Color-coding is based on the original study's assignments.