Supplementary Material for "JIND: Joint Integration and Neural Discrimination for automated single-cell annotation"

Mohit Goyal, Guillermo Serrano, Ilan Shomorony, Idoia Ochoa, Mikel Hernaez

Contents

List of Tables

List of Figures

Additional results

- Table S1 shows the impact of using Seurat Integation prior to training JIND prediction model. Specifically, we compare two cases, Batched: when JIND and JIND+ are evaluated on datasets containing batch effects versus, Integrated: when JIND and JIND+ are run after Seurat integration on the same datasets. We observe that performing integration using Seurat actually hurts the classification performance as compared to directly using JIND or JIND+.
- Figure S2 shows different gene expression patterns across the two groups, G1: *Ductal* cells predicted as *Ductal* and , G2: *Ductal* cells predicted as *Acinar*, allowing differentiation between the two populations of cells. On the contrary, when DE analysis was performed on the groups, G1: randomly chosen subset of *Ductal* cells and , G2: remaining *Ductal* cells, (Figure S4) we neither observe meaningful clustering nor descriptive gene expression patterns necessary for differentiation. We can see the same results on the PBMC dataset on Figure S3 and Figure S5. A summary for this experiment is provided in Figure S6 which shows the differentially expressed gene patterns between accurate predictions and misclassifications made by JIND on cells with a chosen cell-type on both *Pancreas* and *PBMC* dataset.

Datasets		**PBMC 10x_v3-10x_v5**		**Pancreas Bar16-Mu16**		**Pancreas Bar16-Seg16**	
		Batched	Integrated	Batched	Integrated	Batched	Integrated
JIND	*raw*	**0.971**	0.968	**0.958**	0.946	**0.987**	0.946
	rej	0.07	0.06	0.05	0.10	0.05	0.08
	eff	0.986	0.985	0.974	0.979	0.997	0.979
JIND+	*raw*	**0.974**	0.971	0.959	**0.961**	**0.992**	0.961
	rej	0.03	0.03	0.03	0.09	0.02	0.05
	eff	0.985	0.978	0.971	0.980	0.997	0.980

Table S1: Comparing performances of JIND and JIND+ when the source and target batches are integrated with Seurat (Integrated) versus when Seurat integration is not performed (Batched). *raw* is the inital accuracy of the classifier, *rej* is the percentage of cells rejected by the classifier and *eff* is the effective accuracy after rejecting unconfident predictions. Best raw accuracy rates among the two cases (batched or integrated) are **boldfaced**.

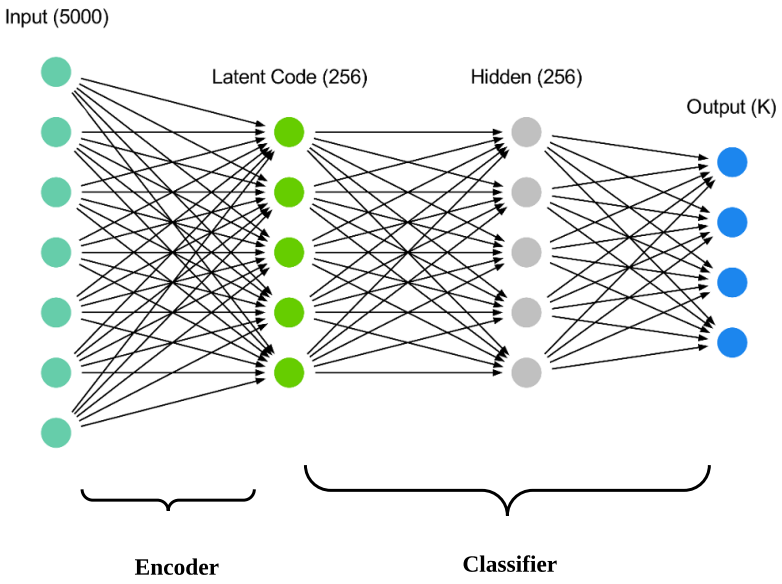

Figure S1: NN-based prediction model employed by JIND for cell-type identification. The network consists of two subnetworks, an encoder and a classifier, which are jointly trained. The encoder subnetwork performs a linear transformation on the gene expression data (of dimension 5000 by default) for a cell and outputs a 256-dimensional latent code. This is then input to the classifier subnetwork which first uses a ReLU activation and then a one hidden-layered (with ReLU activation) classifier outputs a probability vector indicating the likelihood of the cell belonging to each of the K classes.

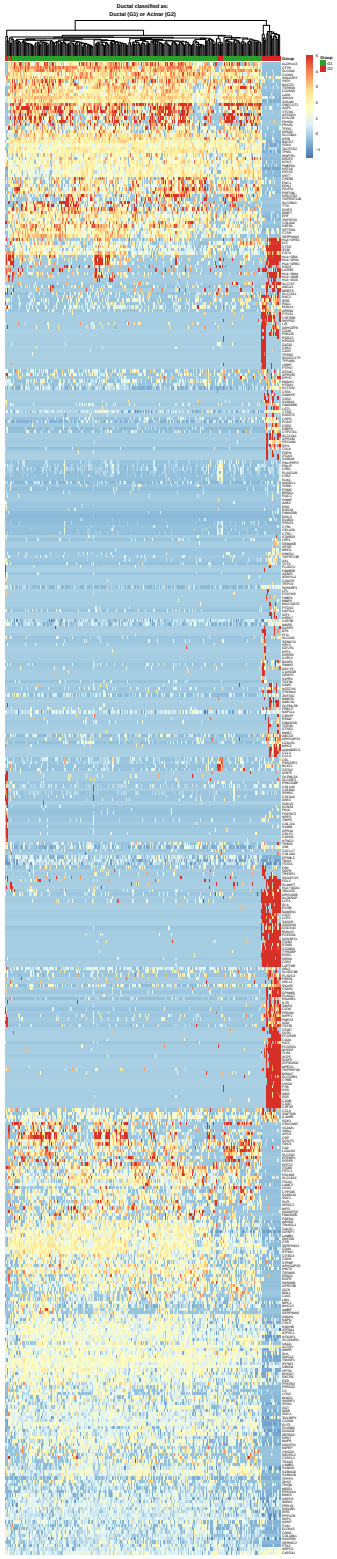

Figure S2: Heatmap for all differentially expressed genes between two groups on *Pancreas* Mur16 dataset. *Ductal* cells predicted by JIND+ as: *Ductal* cells (G1) or *Acinar* cells (G2)

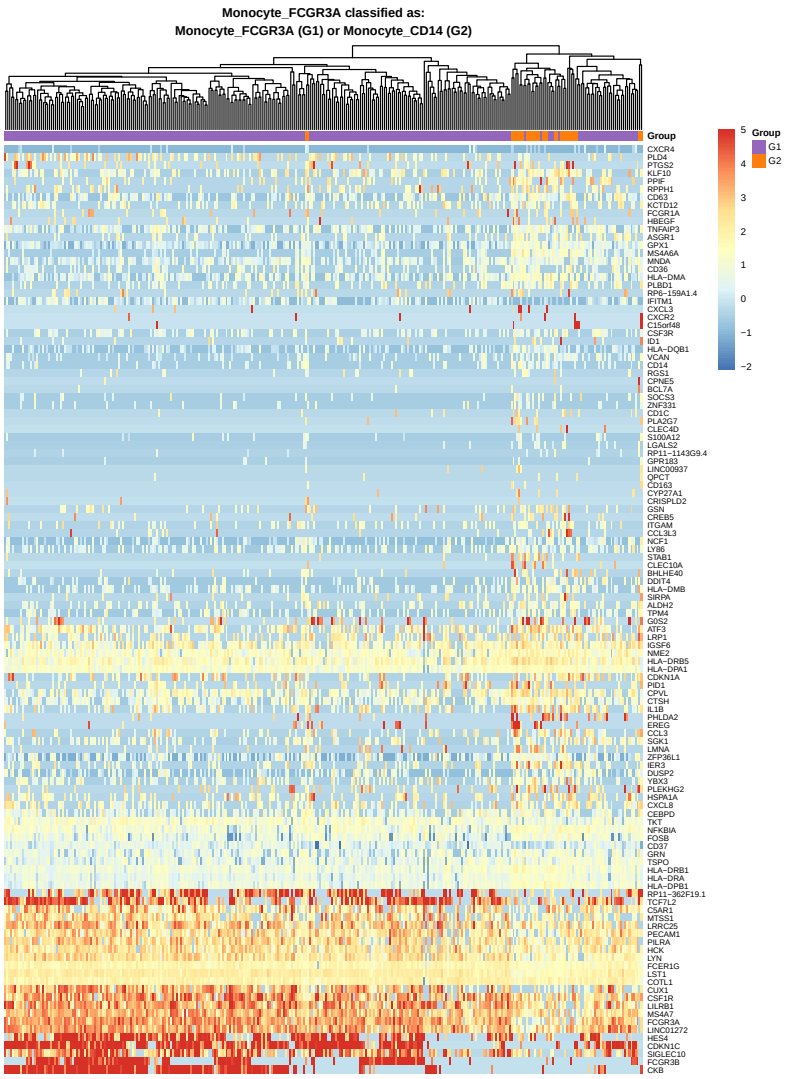

Figure S3: Heatmap for all differentially expressed genes between two groups on *PBMC* 10x_v5 dataset. *Monocytes FCGR3A* cells predicted by JIND+ as: *Monocytes FCGR3A* cells (G1) or *Monocytes CD14* cells (G2).

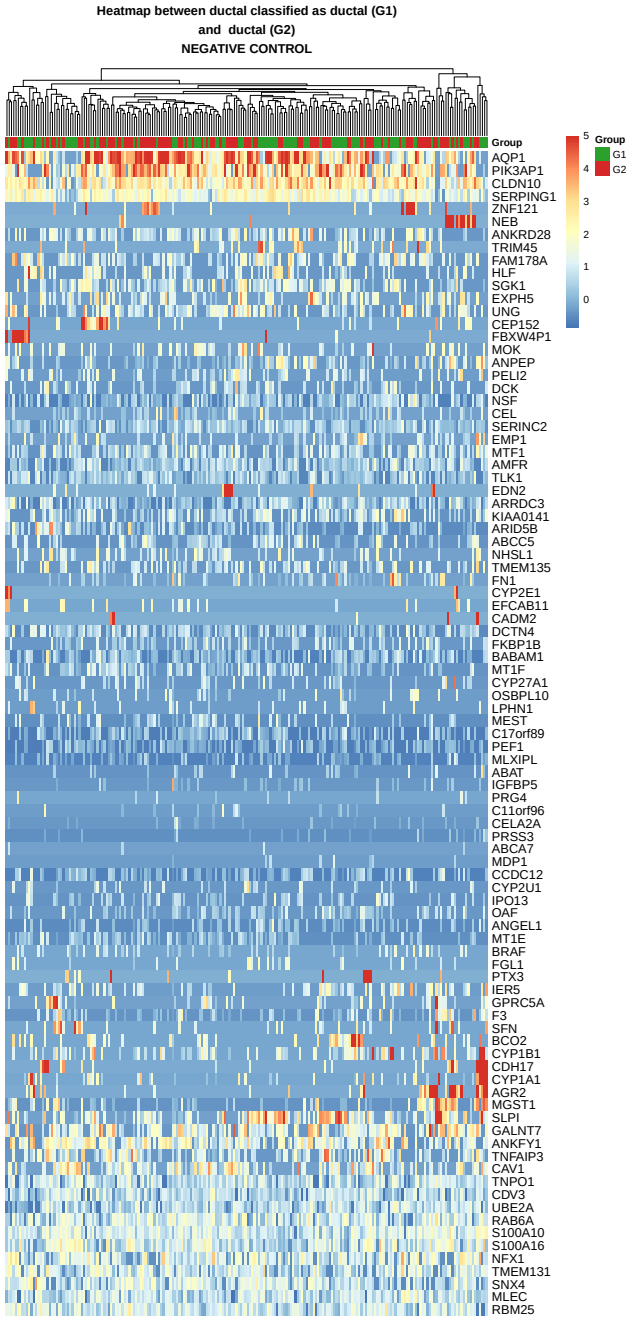

Figure S4: Heatmap for all differentially expressed genes among two randomly chosen groups of *Acinar* cells present in *Pancreas* Mur16 dataset.

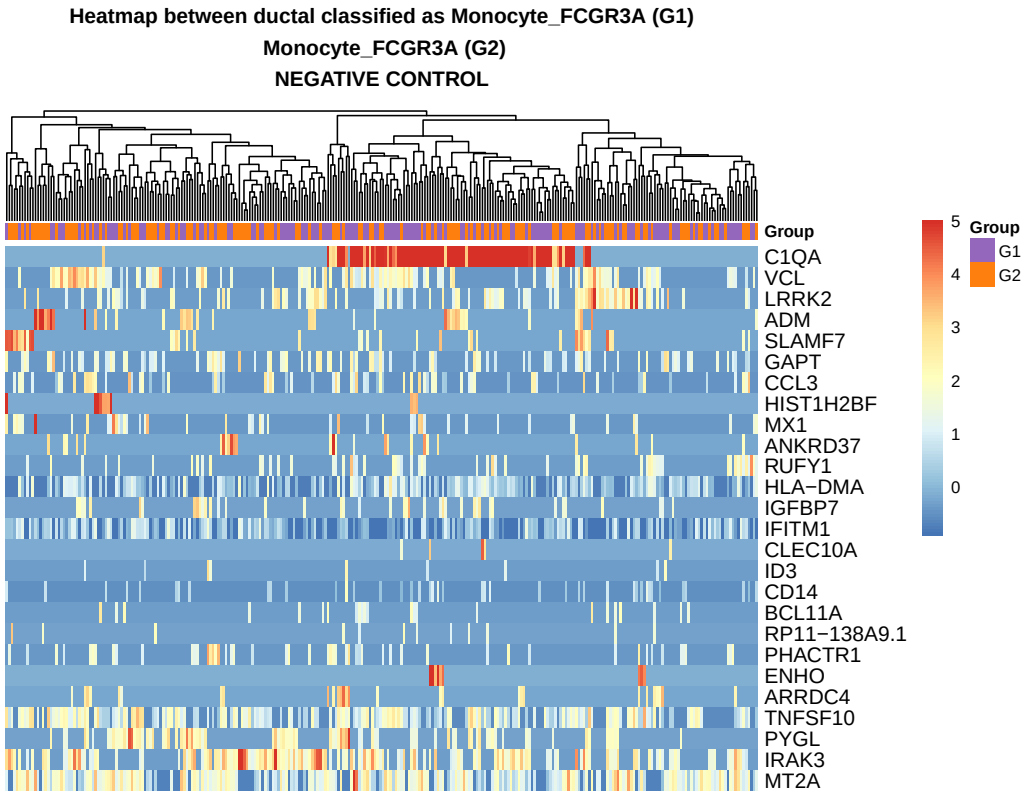

Figure S5: Heatmap for all differentially expressed genes among two randomly chosen *Monocyte FCGR3A* cells present in *PBMC* 10x_v5 dataset.

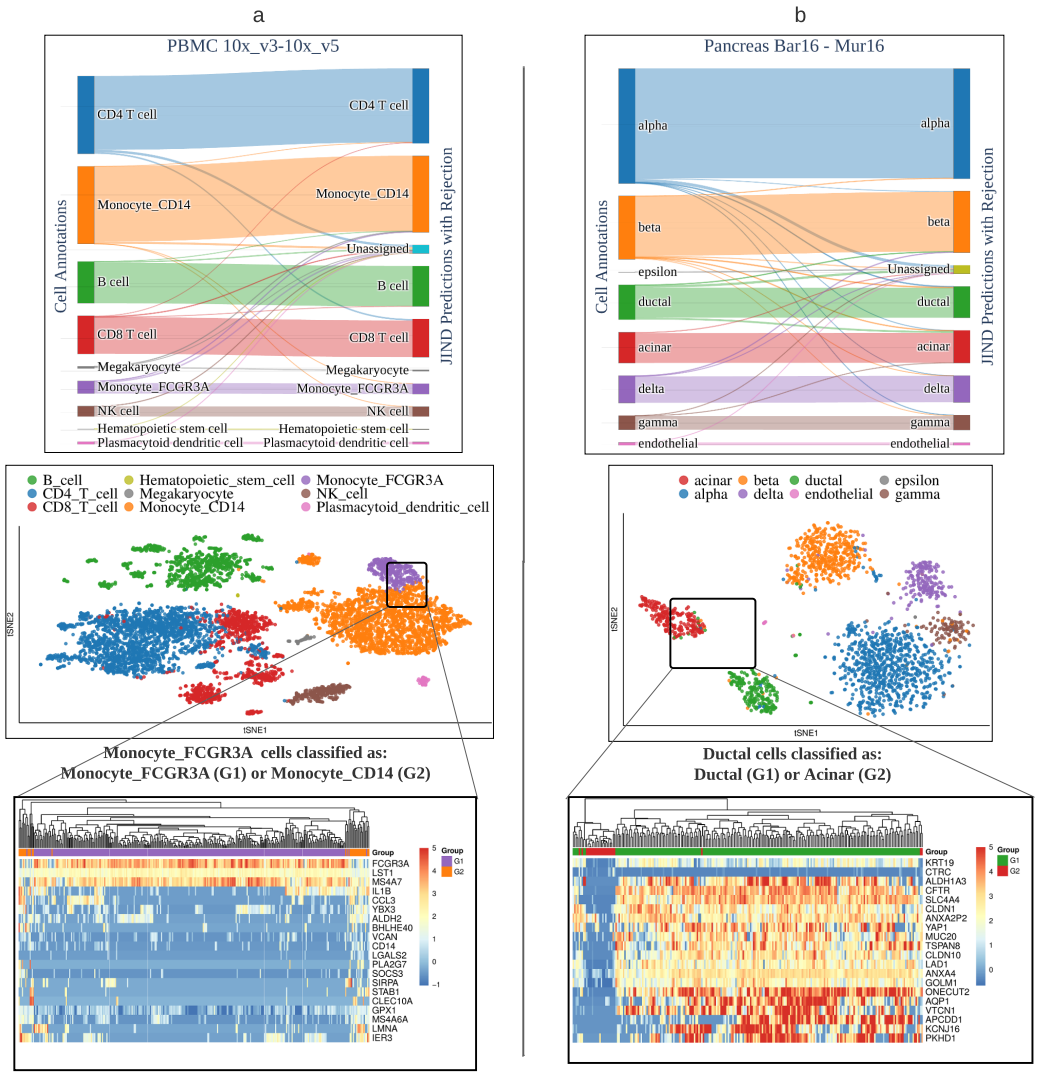

Figure S6: **Performance evaluation and differential expression analysis on two datasets.** The alluvial plots (top) reflect the performance of JIND+ on a) *PBMC* 10x_v3-10x_v5 and b) *Pancreas* Bar16-Mur16 datasets. The tSNE plots (middle) illustrate the cell-type clusters of the target batch, and highlight the two cell-types with the highest misclassification rates: a) *Monocyte_FCGR3A* and *Monocyte_CD14* and b) *Acinar* and *Ductal*. The heatmaps (bottom) show the top 20 differentially expressed genes between a) *Monocyte_FCGR3A* cells classified as *Monocyte_FCGR3A* (G1) and *Monocyte_FCGR3A* classified as *Monocyte_CD14* (G2), and between b) *Ductal* cells classified as *Ductal* (G1) and *Ductal* cells classified as *Acinar* (G2). The shown hierarchical clustering is performed using all the differentially expressed genes.

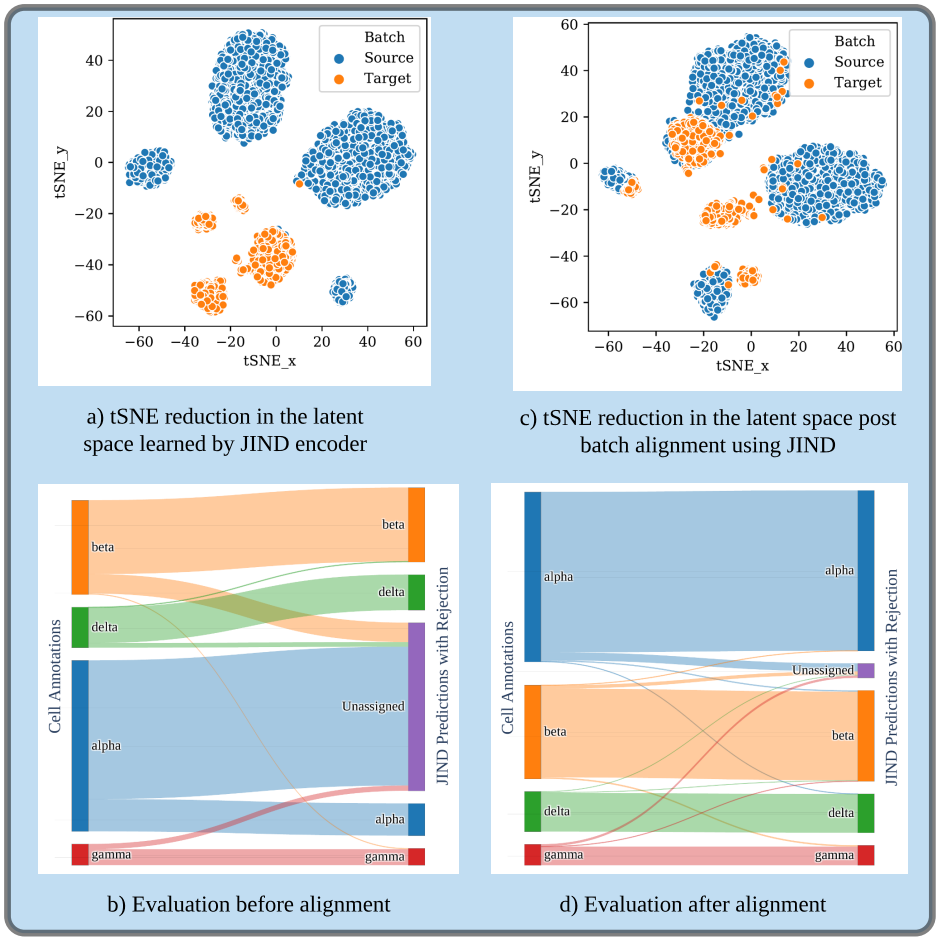

Figure S7: **JIND's asymmetric alignment leads to accurate annotations on batched data.** We consider a subset of cell-types (*Alpha, Beta, Gamma* and *Delta*) from *Pancreas* Bar16 (source batch) and Mur16 (target batch). a) tSNE reduction in the latent space shows significant distributional mismatch due to batch effects. b) As a result, the alluvial plot shows that the prediction model (without alignment) makes a large number of "unassigned" predictions. c) JIND batch alignment removes these batch effects using adversarial training (learning the Generator and Discriminator parameters), which minimizes the distributional discrepancies among the two batches in the latent space learned by the encoder subnetwork. d) The alluvial plot thus obtained after performing batch alignment on target batch shows accurate classification performance per cell-type.