SECANT: a biology-guided semi-supervised method for clustering, classification, and annotation of single-cell multi-omics

Xinjun Wang[1], Zhongli Xu[2], Xueping Zhou[1], Yanfu Zhang[3], Robert Lafyatis[4], Kong Chen[4], Heng Huang[3], Ying Ding[1], Richard H. Duerr[4], and Wei Chen[1,2,#]
Department of [1] Biostatistics; [2] Pediatrics; [3] Electrical and Computer Engineering; [4] Medicine
University of Pittsburgh, Pittsburgh, PA15261

BACKGROUND

Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-seq)

- One of the first and most popular techniques that can measure surface protein and mRNA expression level simultaneously in the same cell
- Oligonucleotide-labeled antibodies are used to integrate protein and transcriptome measurements into an efficient, single-cell readout (**Fig 1**)
- A CITE-Seq panel of dozens well-characterized monoclonal antibodies that recognize cell-surface proteins are routinely used as markers (**Fig 2**)

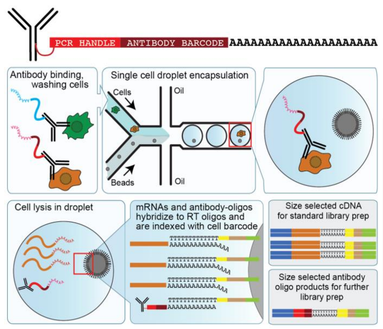


Fig 1. An overview of droplet-based CITE-seq techniques by *10X Genomics* protocol (picture from *https://cite-seq.com/*)

	Transcriptomic mRNA Data					Proteomic ADT Data				
	gene1	gene2	gene3	…	gene1000	CD14	CD16	CD19	CD56	CD8
cell1	10	0	3	…	72	36	59	1127	25	104
cell2	5	0	18	…	51	82	16	68	43	196
…	…	…	…	…	…	…	…	…	…	…
cell1000	0	2	41	…	4	38	19	76	1105	500
cell1001	6	3	15	…	21	47	53	1296	53	128
…	…	…	…	…	…	…	…	…	…	…
cell10000	0	5	10	…	39	53	12	75	1359	251

Fig 2. UMI count matrices of paired transcriptomic and proteomic data

Analyzing Transcriptomic and Proteomic Data Individually (Fig 3)

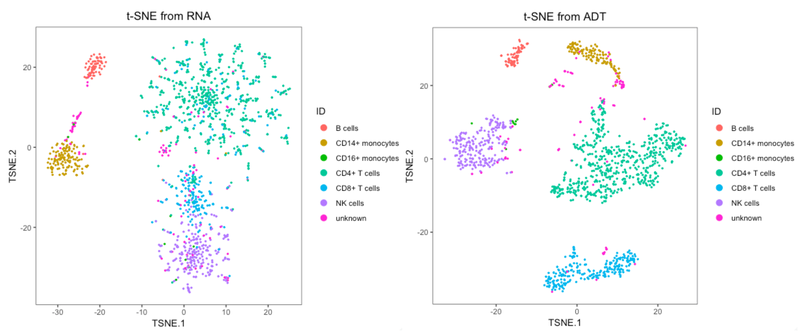


Fig 3. t-SNE plots based on RNA (left) and protein data (right) to identify immune cell types in PBMCs

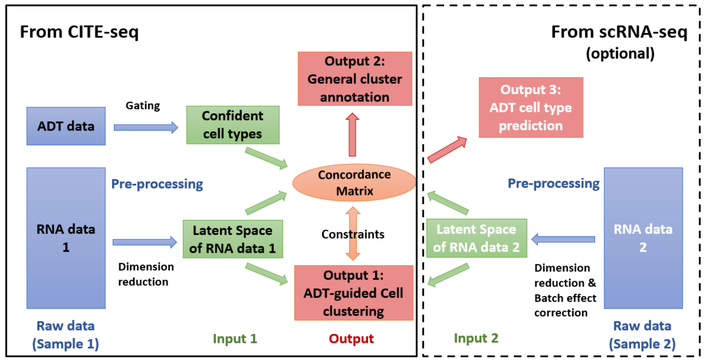
MOTIVATIONS

- Novel methods for single cell multi-omics (e.g., CITE-seq) are in urgent need
- Existing methods utilize data-driven approach but none of them incorporate existing biological knowledge (e.g., from flow/mass cytometry)
- Novelties of SECANT include:
  - Using confident cell types label identified from surface protein data as guidance for cell clustering with RNA data
  - Providing general annotation of confident cell types for each cell cluster
  - Fully utilizing cells with uncertain or missing cell type label (e.g., as a result from gating)
  - Jointly analyzing CITE-seq and scRNA-seq data to predict confident cell types label for scRNA-seq data
  - A model-based approach to provide clustering and classification uncertainty

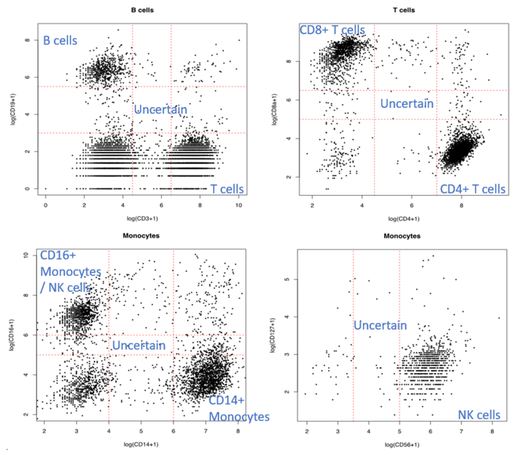
Contact Information: Xinjun Wang, xiw119@pitt.edu

METHODS

General workflow of SECANT



An example of gating with surface protein data from CITE-seq



Core assumption of SECANT

- Cells shouldn't fall into the same cluster from RNA data if they are identified as different confident cell types with surface protein data

		Clusters from RNA data						
		Cluster *1*	Cluster *2*	…	Cluster *k*	Cluster *k+1*	…	Cluster *K*
Confident cell types from ADT data	Cell Type *1*	p_{11}	0		0	*0*		0
	Cell Type *2*	0	p_{22}		0	*0*		0
	…							
	Cell Type *m*	0	0		p_{mk}	$p_{m,k+1}$		0
	…							
	Cell Type *M*	0	0		0	*0*		p_{MK}
	Uncertain*	1 - p_{11}	1 - p_{22}		1 - p_{mk}	1 – $p_{m,k+1}$		1 - p_{MK}

Optimizing likelihood though stochastic gradient descent (SGD)

- Y_{ij}: the observed latent space of RNA data (e.g., from scVI)
  - i: cell index; j: feature index
- L_i: the observed confident cell type label from ADT data for cell i
- Z_i: the hidden true cluster label for cell i
- K: the total number of clusters
- $P(Y,L) = \prod_{i=1}^{N}\{\sum_{k=1}^{K} P(L_i|Z_i=k)P(Z_i=k|Y_i)\}\prod_{i=1}^{N}\{\prod_{k=1}^{K} f(Y_i|\theta_k)^{1(Z_i=k)}\}$

REAL DATA APPLICATION

Analysis of a public human PBMC sample

- 7,865 cells and 14 surface protein markers (from *10X Genomics*)

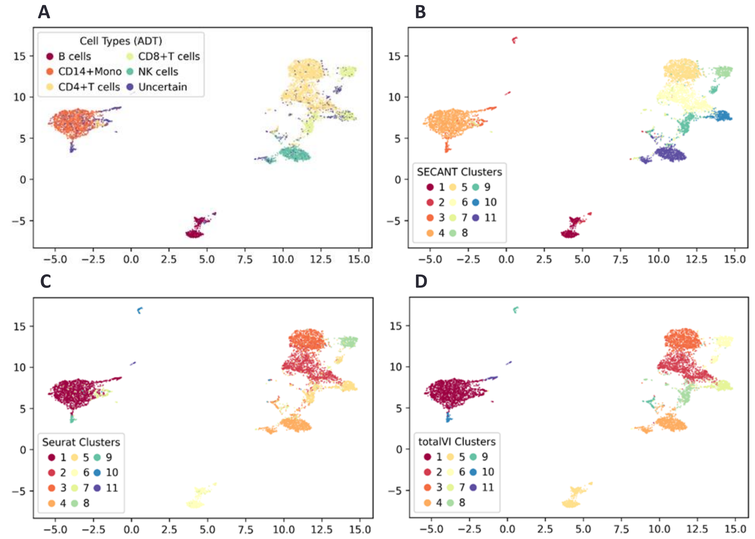


Fig 4. UMAP visualization of the latent space of RNA data. A: ADT confident cell types built with manual gating. B: SECANT result. C: Seurat result. D: colored by totalVI result

Table 1. Estimated concordance matrix and post-hoc subtype identification from SECANT

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10	Cluster 11
B cells	0.969	0.474	0	0	0	0	0	0	0	0	0
CD14+ Monocytes	0	0	0.203	0.877	0	0	0	0	0	0	0
CD4+ T cells	0	0	0	0	0.986	0.911	0.627	0	0	0	0
CD8+ T cells	0	0	0	0	0	0	0	0.939	0.741	0.745	0
NK Cells	0	0	0	0	0	0	0	0	0	0	0.884
Uncertain	0.031	0.526	0.797	0.123	0.014	0.089	0.373	0.061	0.259	0.255	0.116
Cluster Weight	0.064	0.031	0.054	0.219	0.152	0.154	0.03	0.049	0.079	0.053	0.117
SECANT Annotation	Follicular B cells	Marginal zone B cells	Dendritic cells	CD14+ Monocytes	Naïve CD4+ T cells	Memory CD4+ T cells	Gamma delta T cells	Naïve CD8+ T cells	Effector CD8+ T cells	Memory CD8+ T cells	NK cells
Select DE Genes	IGHM CD79A MS4A1 IGHD CD22	MZB1 TNFRSF17 CD1 CD27	CSF1R CST3	S100A9 S100A8 LYZ	CCR7 SELL CD3D	TRAC IL7R LTB	GZMK KLRB1	SELL CCR7	CD8B CD8A GZMK NKG7 GZMM	GZMK IL7R CD8A CD69 KLRB1	

CONCLUSIONS

- SECANT will be complementary to existing tools for characterizing novel cell types and make new biological discoveries
- Our program (in PyTorch) runs ~ 40X faster on GPU than CPU, and most of the analyses were conducted using the GPU resources from Pitt CRC
- Our manuscript has been revision submitted to ***Genome Research***

ACKNOWLEDGMENTS

- This work was funded by National Institutes of Health grants [P01AI106684 to W.C., U01DK062420 to W.C., R.D., P50AR060780 to R.L. and W.C., R01HL137709 to K.C.], and was also supported in part by Children's Hospital of Pittsburgh of the UPMC Health System, and **the University of Pittsburgh Center for Research Computing through the resources provided**.

BACKGROUND

Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-seq)

- One of the first and most popular techniques that can measure surface protein and mRNA expression level simultaneously in the same cell
- Oligonucleotide-labeled antibodies are used to integrate protein and transcriptome measurements into an efficient, single-cell readout (**Fig 1**)
- A CITE-Seq panel of dozens well-characterized monoclonal antibodies that recognize cell-surface proteins are routinely used as markers (**Fig 2**)

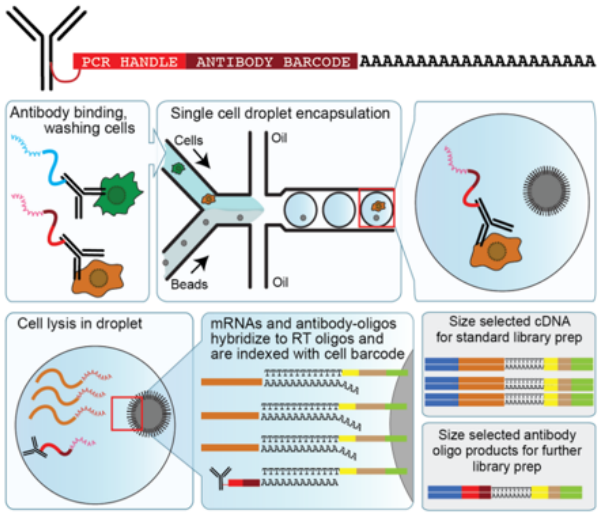


Fig 1. An overview of droplet-based CITE-seq techniques by *10X Genomics* protocol (picture from *https://cite-seq.com/*)

	Transcriptomic mRNA Data					Proteomic ADT Data				
	gene1	gene2	gene3	...	gene1000	CD14	CD16	CD19	CD56	CD8
cell1	10	0	3	...	72	36	59	1127	25	104
cell2	5	0	18	...	51	82	16	68	43	196
...	...	...	...	...	...	...	...	...	...	...
cell1000	0	2	41	...	4	38	19	76	1105	500
cell1001	6	3	15	...	21	47	53	1296	53	128
...	...	...	...	...	...	...	...	...	...	...
cell10000	0	5	10	...	39	53	12	75	1359	251

Fig 2. UMI count matrices of paired transcriptomic and proteomic data

Analyzing Transcriptomic and Proteomic Data Individually (Fig 3)

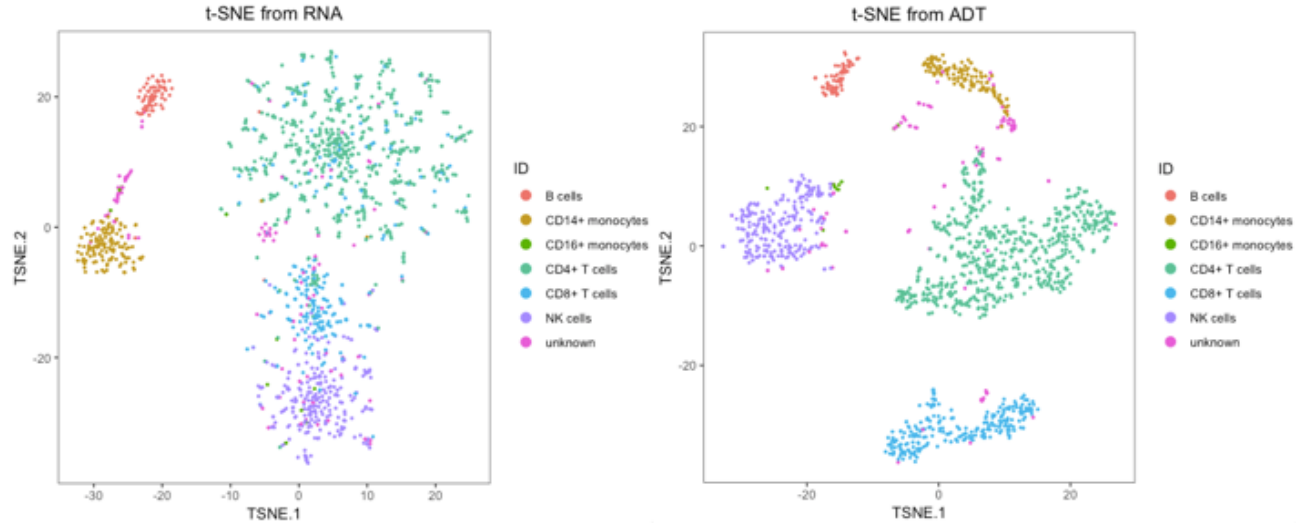


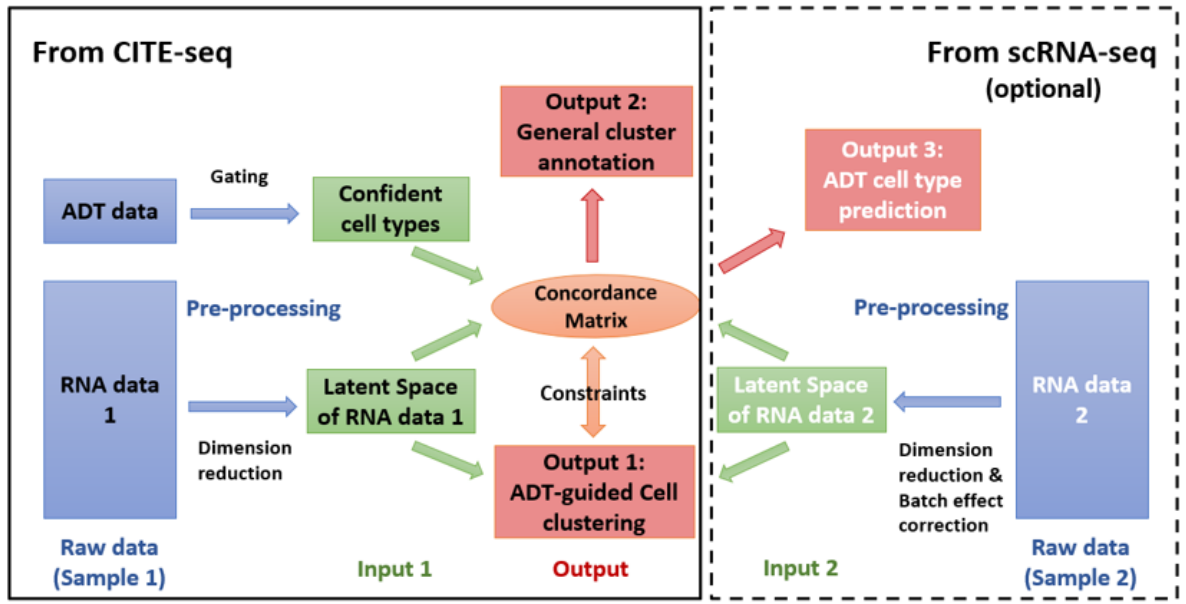
Fig 3. t-SNE plots based on RNA (left) and protein data (right) to identify immune cell types in PBMCs

MOTIVATIONS

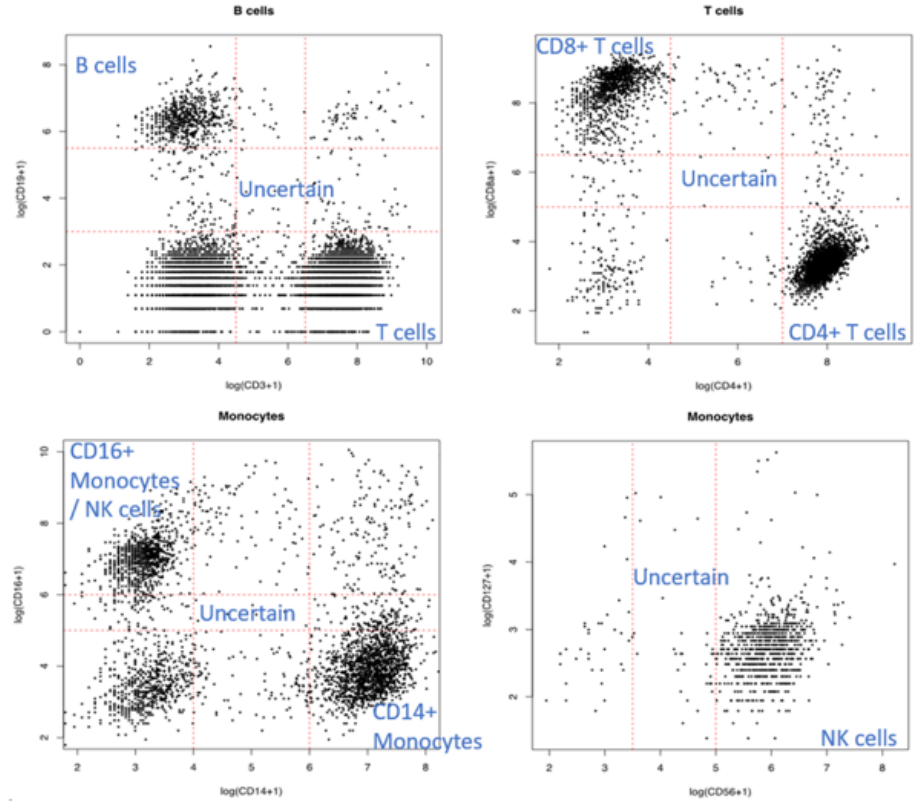
- Novel methods for single cell multi-omics (e.g., CITE-seq) are in urgent need
- Existing methods utilize data-driven approach but none of them incorporate existing biological knowledge (e.g., from flow/mass cytometry)
- Novelties of SECANT include:
  - Using confident cell types label identified from surface protein data as guidance for cell clustering with RNA data
  - Providing general annotation of confident cell types for each cell cluster
  - Fully utilizing cells with uncertain or missing cell type label (e.g., as a result from gating)
  - Jointly analyzing CITE-seq and scRNA-seq data to predict confident cell types label for scRNA-seq data
  - A model-based approach to provide clustering and classification uncertainty

METHODS

General workflow of SECANT



An example of gating with surface protein data from CITE-seq



Core assumption of SECANT

- Cells shouldn't fall into the same cluster from RNA data if they are identified as different confident cell types with surface protein data

		Clusters from RNA data						
		Cluster *1*	Cluster *2*	...	Cluster *k*	Cluster *k+1*	...	Cluster *K*
Confident cell types from ADT data	Cell Type *1*	p_{11}	0		0	*0*		0
	Cell Type *2*	0	p_{22}		0	*0*		0
	...							
	Cell Type *m*	0	0		p_{mk}	$p_{m,k+1}$		0
	...							
	Cell Type *M*	0	0		0	*0*		p_{MK}
	Uncertain*	1 - p_{11}	1 - p_{22}		1 - p_{mk}	1 – $p_{m,k+1}$		1 - p_{MK}

Optimizing likelihood though stochastic gradient descent (SGD)

- Y_{ij}: the observed latent space of RNA data (e.g., from scVI)
  - i: cell index; j: feature index
- L_i: the observed confident cell type label from ADT data for cell i
- Z_i: the hidden true cluster label for cell i
- K: the total number of clusters
- $P(Y, L) = \prod_{i=1}^{N}\{\sum_{k=1}^{K} P(L_i|Z_i = k)P(Z_i = k|Y_i)\}\prod_{i=1}^{N}\{\prod_{k=1}^{K} f(Y_i|\theta_k)^{1(Z_i=k)}\}$

REAL DATA APPLICATION

Analysis of a public human PBMC sample

- 7,865 cells and 14 surface protein markers (from *10X Genomics*)

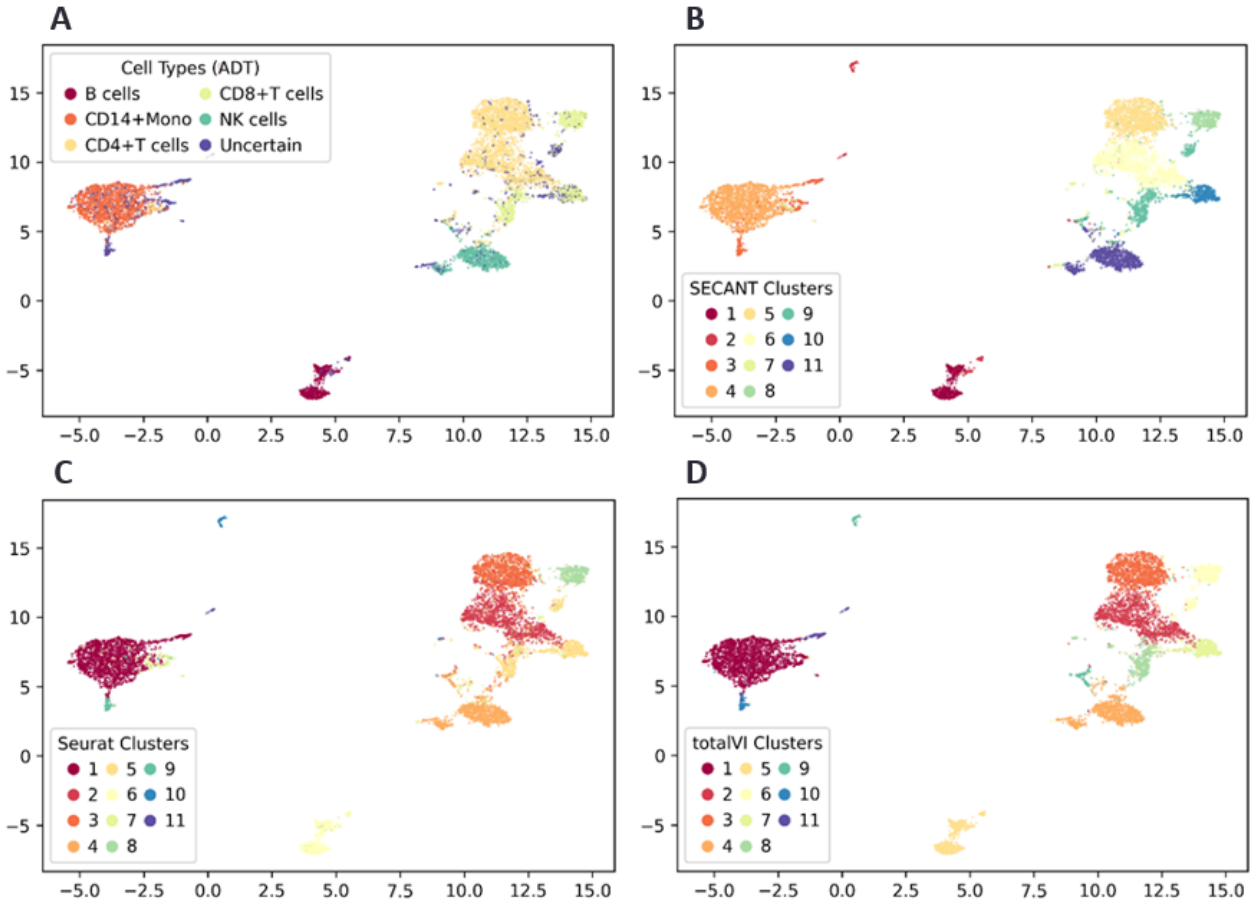


Fig 4. UMAP visualization of the latent space of RNA data. A: ADT confident cell types built with manual gating. B: SECANT result. C: Seurat result. D: colored by totalVI result

Table 1. Estimated concordance matrix and post-hoc subtype identification from SECANT

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10	Cluster 11
B cells	0.969	0.474	0	0	0	0	0	0	0	0	0
CD14+ Monocytes	0	0	0.203	0.877	0	0	0	0	0	0	0
CD4+ T cells	0	0	0	0	0.986	0.911	0.627	0	0	0	0
CD8+ T cells	0	0	0	0	0	0	0	0.939	0.741	0.745	0
NK Cells	0	0	0	0	0	0	0	0	0	0	0.884
Uncertain	0.031	0.526	0.797	0.123	0.014	0.089	0.373	0.061	0.259	0.255	0.116
Cluster Weight	0.064	0.031	0.054	0.219	0.152	0.154	0.03	0.049	0.079	0.053	0.117
SECANT Annotation	Follicular B cells	Marginal zone B cells	Dendritic cells	CD14+ Monocytes	Naïve CD4+ T cells	Memory CD4+ T cells	Gamma delta T cells	Naïve CD8+ T cells	Effector CD8+ T cells	Memory CD8+ T cells	NK cells
Select DE Genes	IGHM CD79A MS4A1 IGHD CD22	MZB1 TNFRSF17 CD1 CD27	CSF1R CST3	S100A9 S100A8 LYZ	CCR7 SELL CD3D	TRAC IL7R LTB	GZMK KLRB1	SELL CCR7	CD8B CD8A GZMK NKG7 GZMM	GZMK IL7R CD8A CD69 KLRB1	

CONCLUSIONS

- SECANT will be complementary to existing tools for characterizing novel cell types and make new biological discoveries
- Our program (in PyTorch) runs ~ 40X faster on GPU than CPU, and most of the analyses were conducted using the GPU resources from Pitt CRC
- Our manuscript has been revision submitted to ***Genome Research***

ACKNOWLEDGMENTS

- This work was funded by National Institutes of Health grants [P01AI106684 to W.C., U01DK062420 to W.C., R.D., P50AR060780 to R.L. and W.C., R01HL137709 to K.C.], and was also supported in part by Children's Hospital of Pittsburgh of the UPMC Health System, and **the University of Pittsburgh Center for Research Computing through the resources provided**.