Studying Cellular Composition Profiles using Grade of Membership Models

Kushal K. Dey

Raphael Gottardo Lab Summer Project 2017

PhD Advisor: Matthew Stephens

Zheng et al (2017) data analysis

10X Genomics team has collected large scale single cell RNA-seq data from a large number of cell types and mix of cell types, along with many unsorted cells.

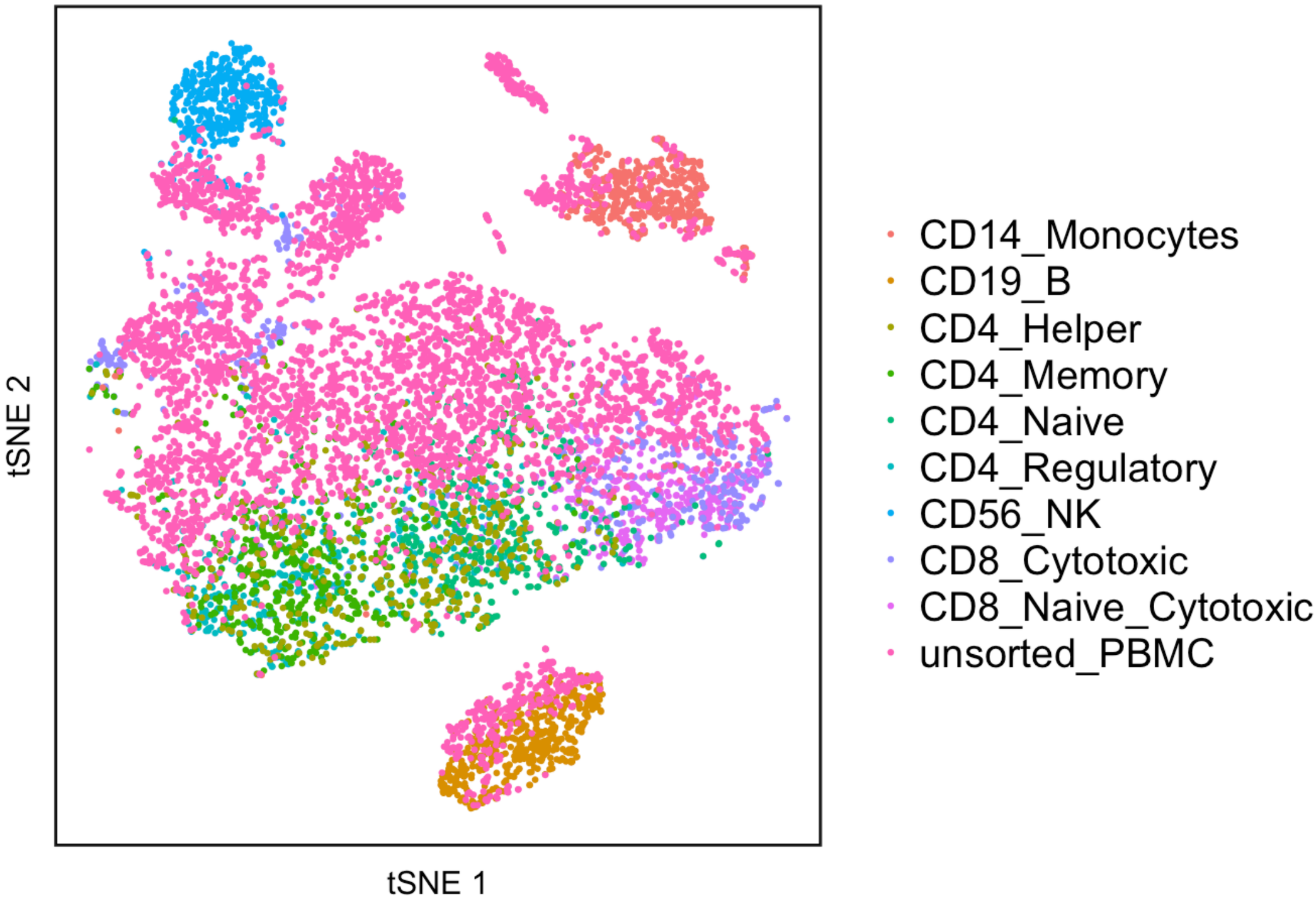
https://support.10xgenomics.com/single-cell-gene-expression/datasets

We selected 500 cells randomly from 9 different sorted cell types.

- ***CD14 monocytes***
- ***CD19 B***
- ***CD4 helper T cells***
- ***CD4 Regulatory T cells***
- ***CD4 Naive T cells***
- ***CD4 Regulatory T cells***
- ***CD56 NK cells***
- ***CD8 Cytotoxic T cells***
- ***CD8 Cytotoxic naive T cells***

We combined this with 5000 randomly selected PBMC cells from the 68K unsorted PBMC data.

t-SNE plot with sorted cells colored



Supervised Learning of unsorted cells based on sorted immune cell types

Comparison of classification methods on sorted immune cells

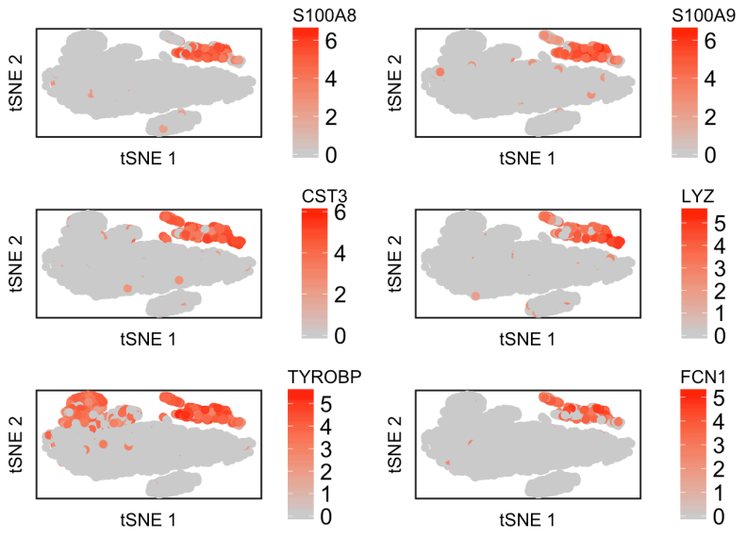
supervised Countclust vs SVM - *comparison 1*

training subsampling rate	supervised CountClust (all genes)	SVM (PC data - top 100 PCs)
0.25	0.24	0.35
0.50	0.2	0.32
0.75	0.2	0.31

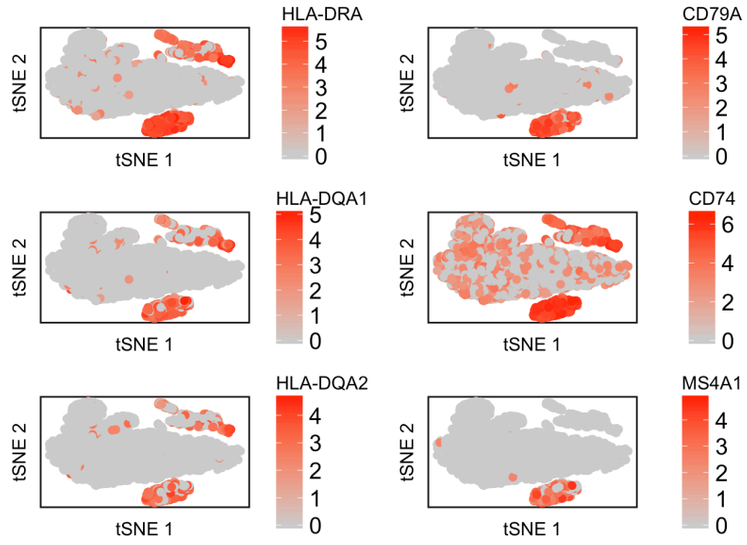
supervised Countclust vs SVM - *comparison 2*

training subsampling rate	supervised CountClust (top varying genes)	SVM(before norm: top varying genes)	SVM (after norm: top varying genes)
0.25	0.35	0.43	0.32
0.50	0.34	0.41	0.31
0.75	0.34	0.40	0.31

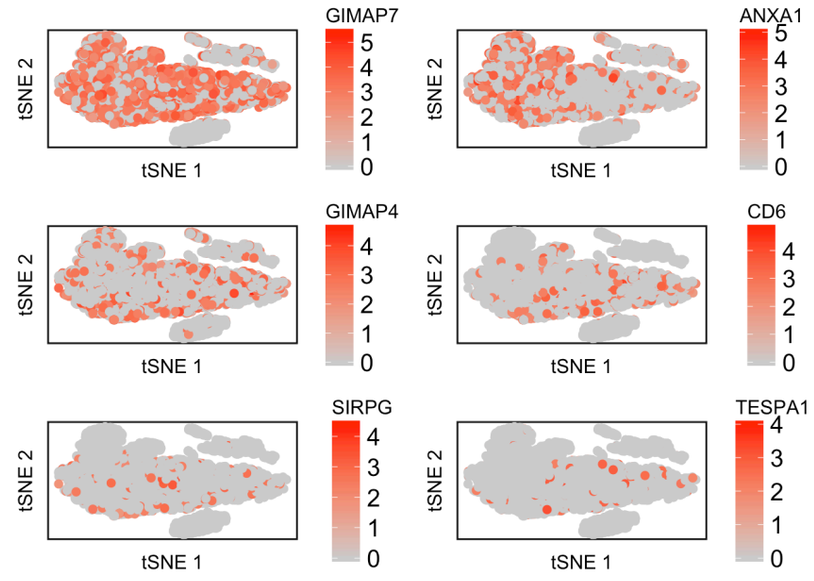
CD14 monocytes



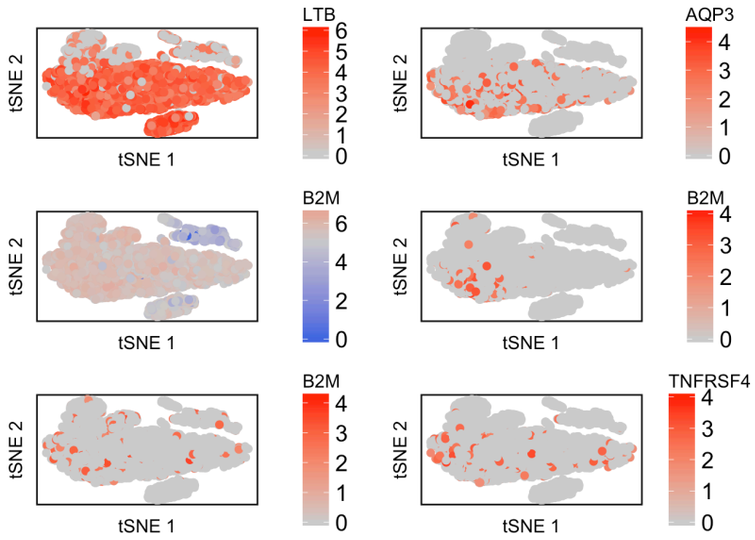
CD19 B



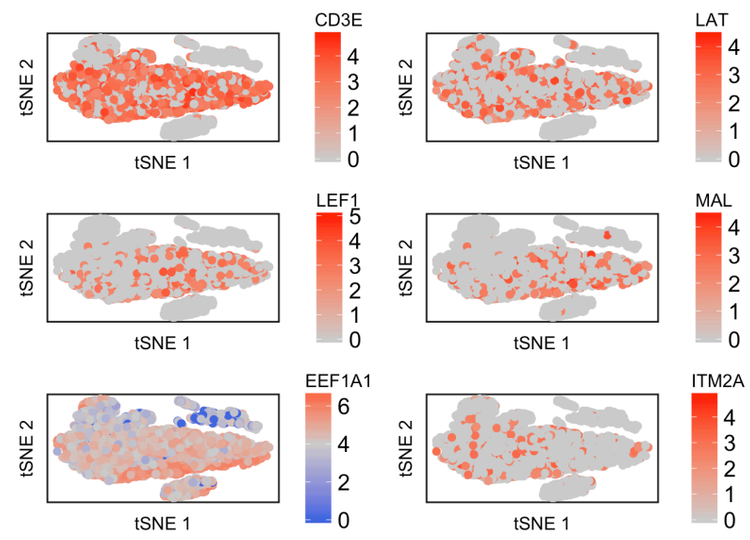
CD4 helper



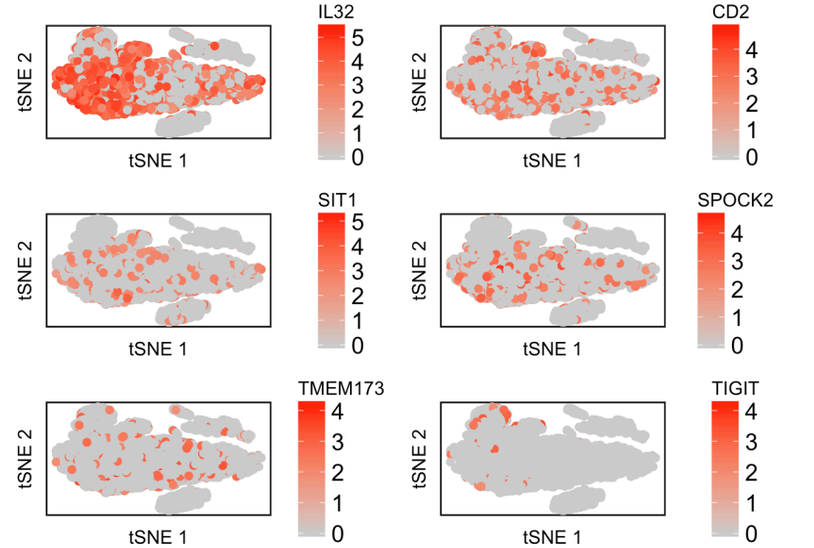
CD4 memory



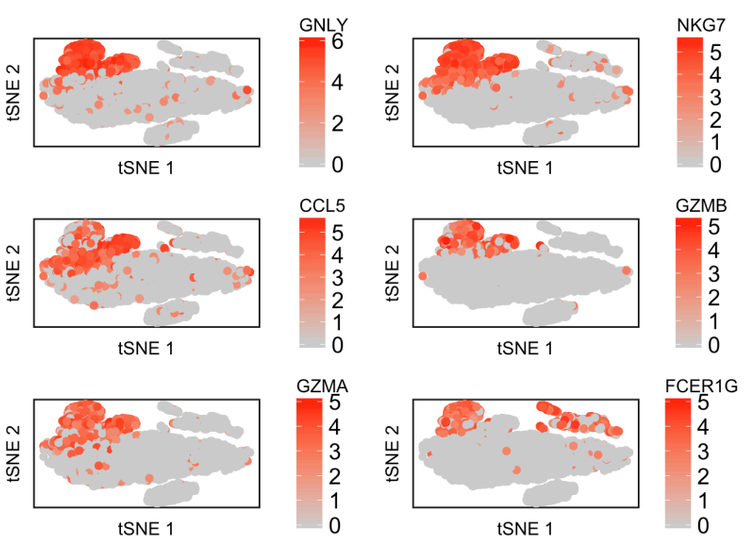
CD4 naive



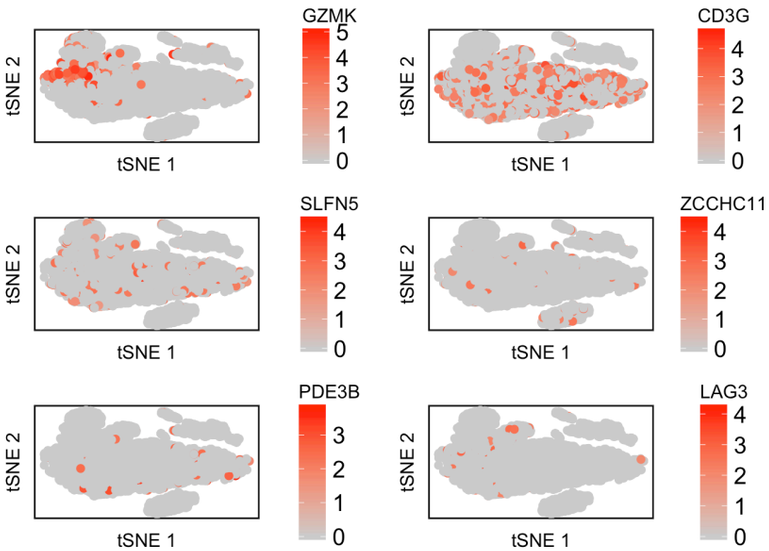
CD4 regulatory



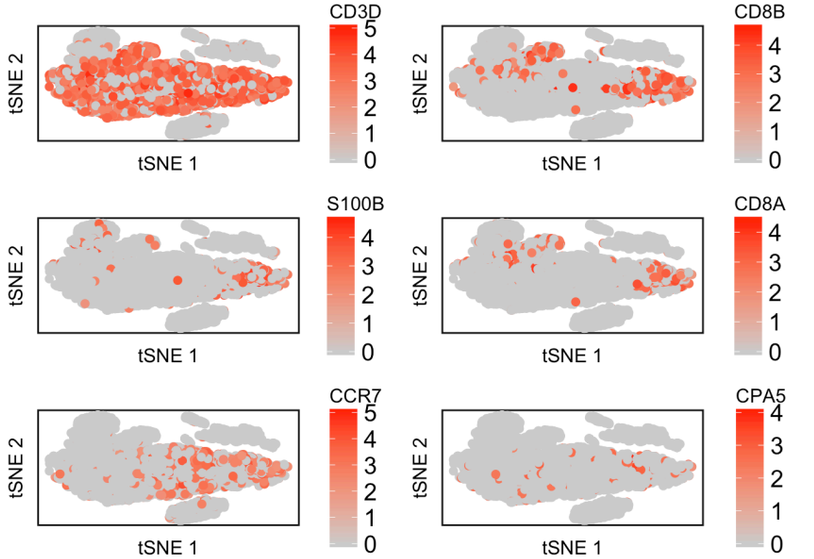
CD56 NK cells



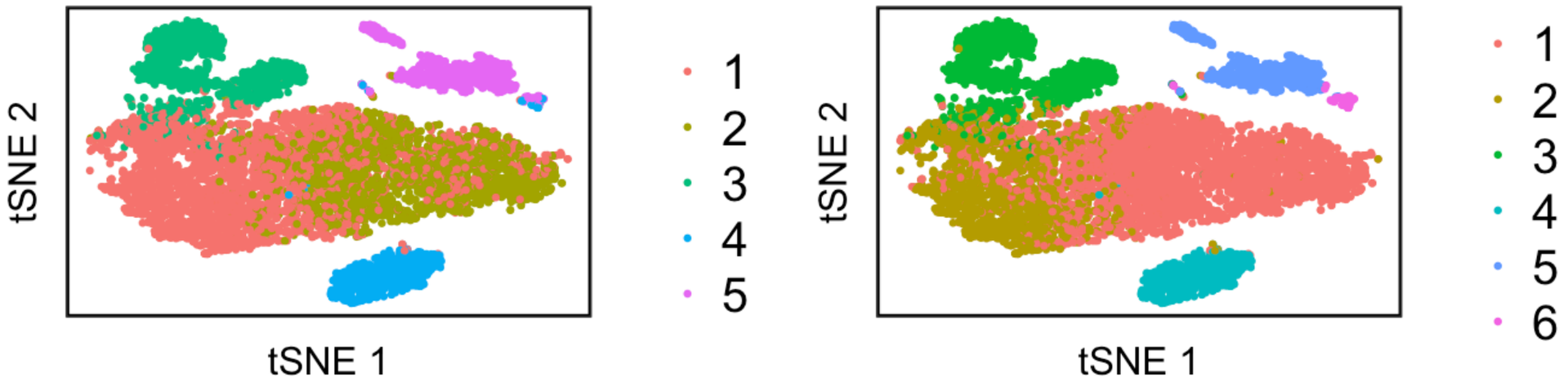
CD8 Cytotoxic

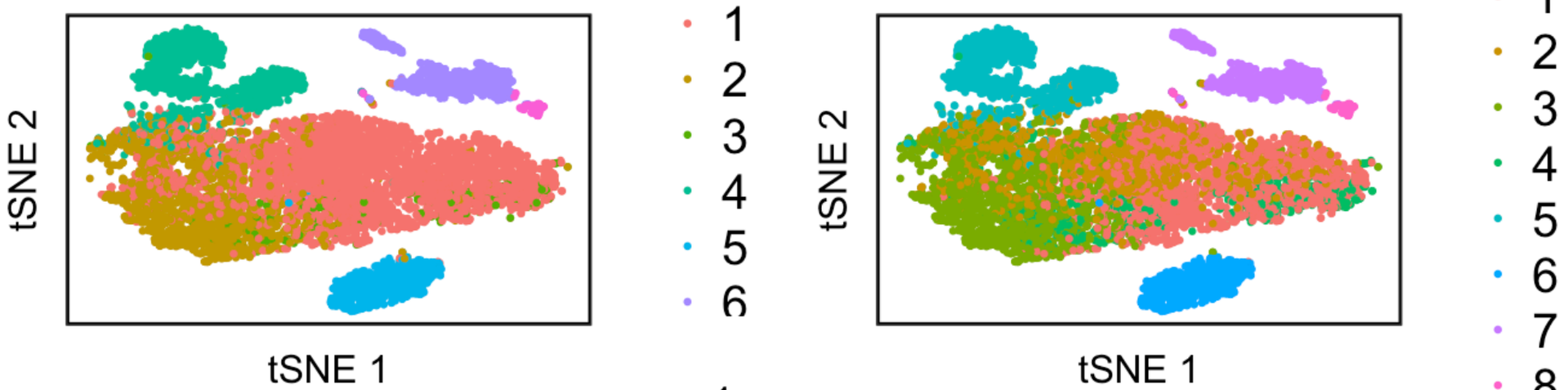


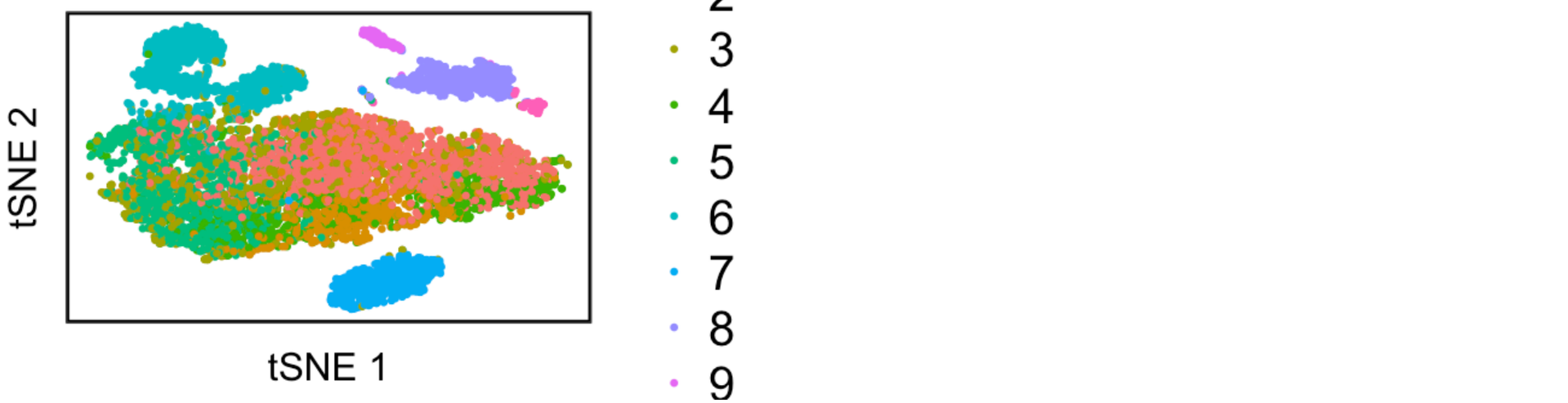
CD8 Naive Cytotoxic



Unsupervised CountClust (range of K values)

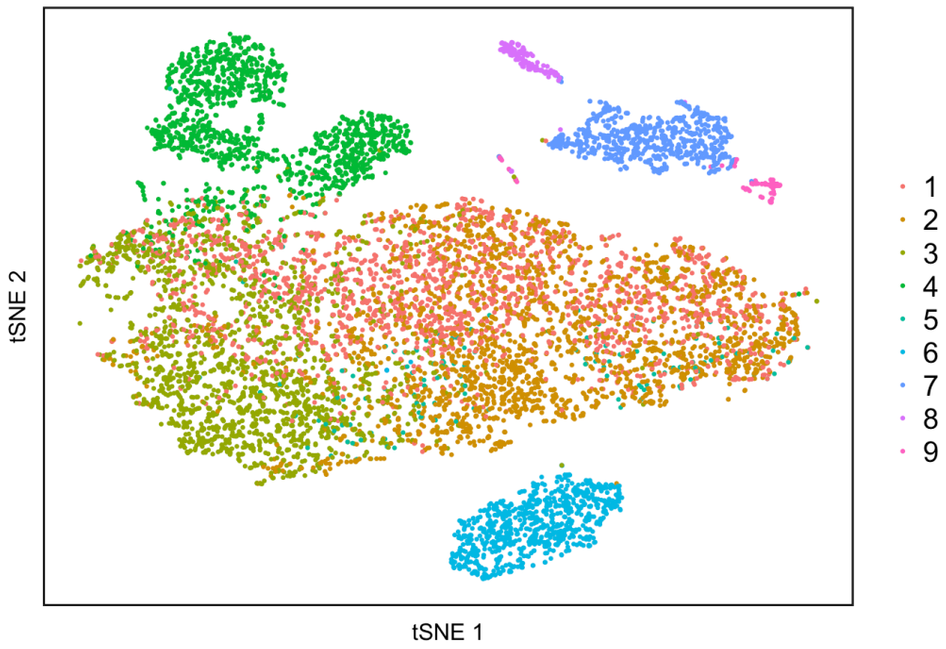




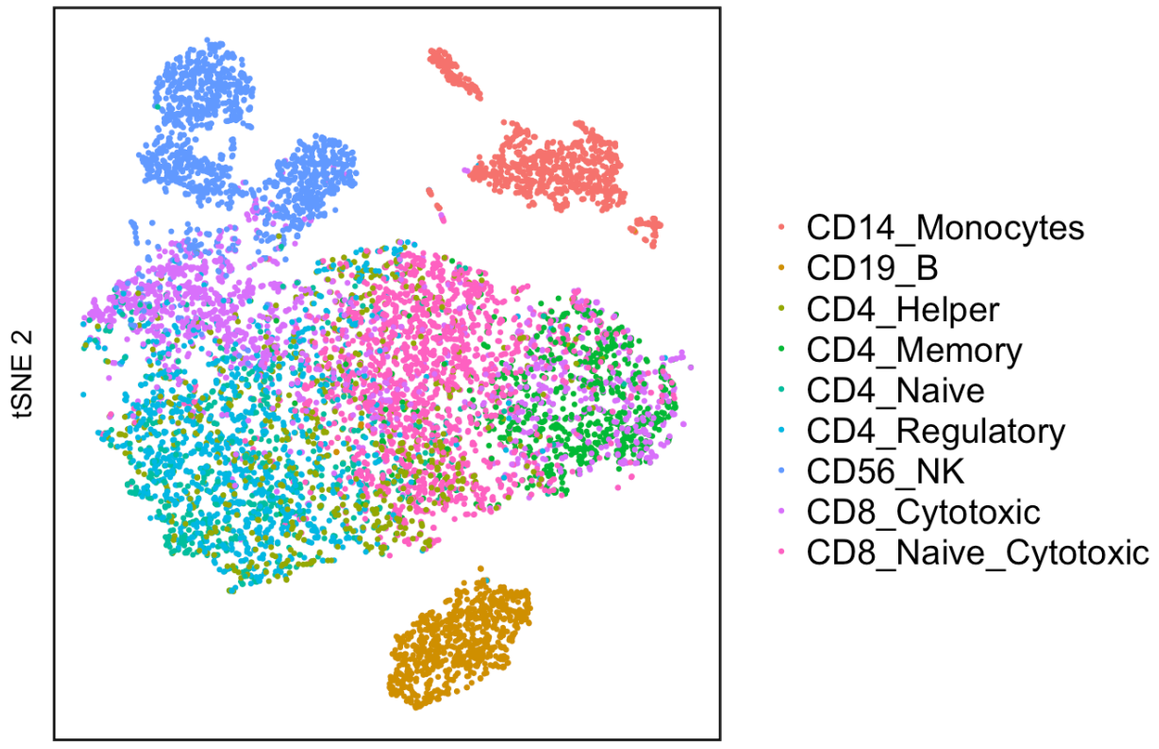


Unsupervised CountClust

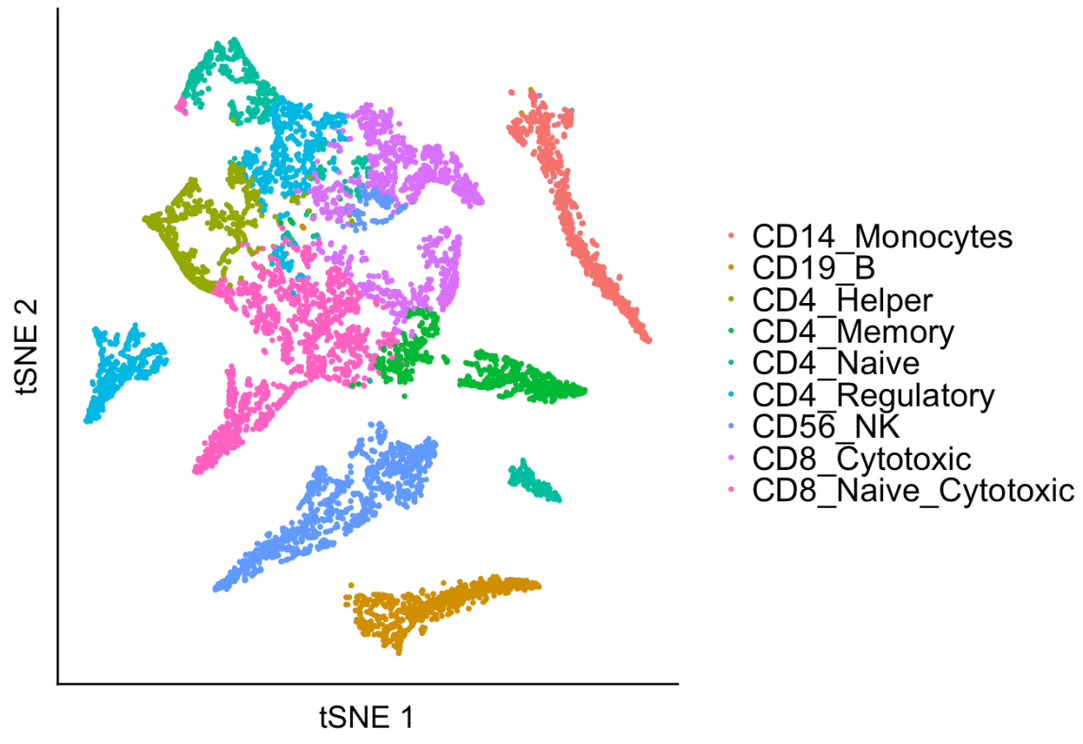
K=9



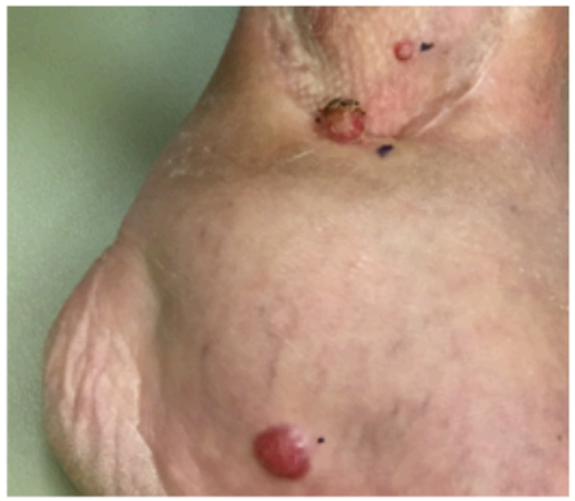
Supervised CountClust



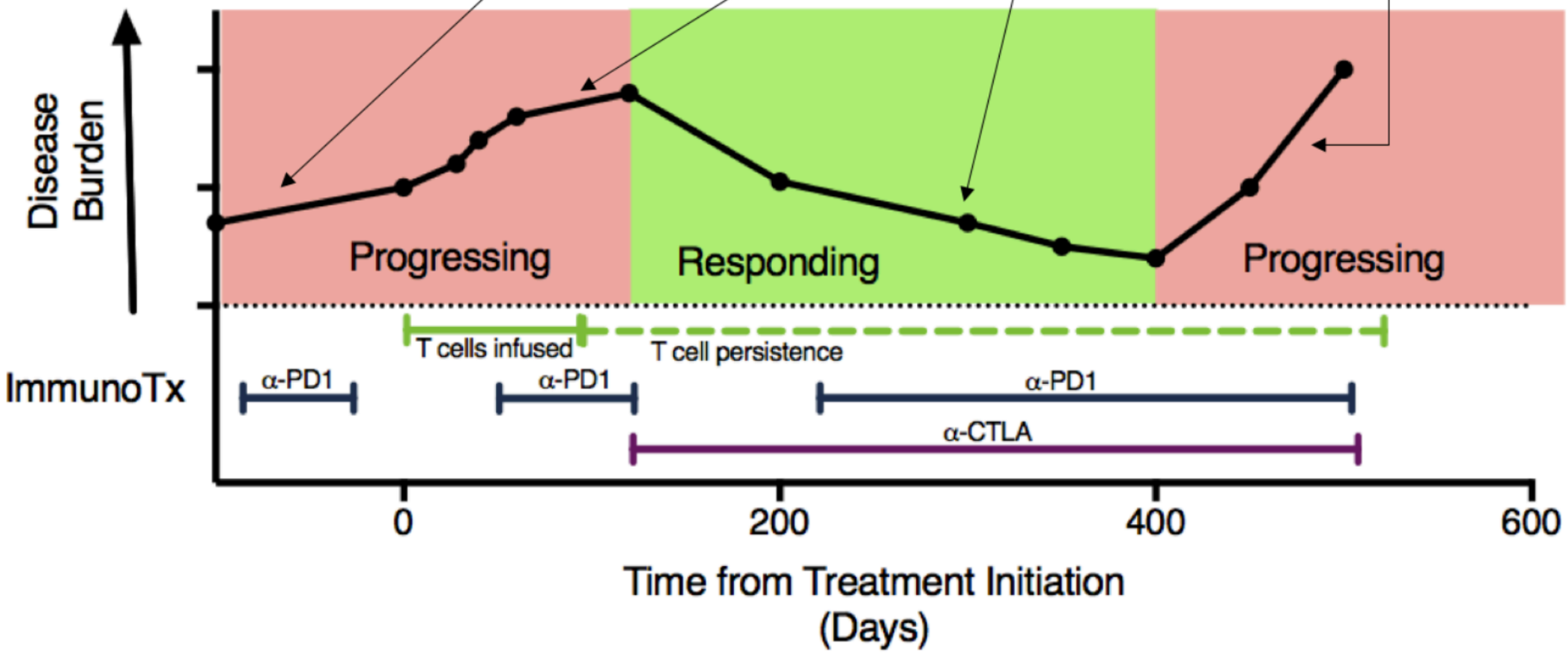
classtpx tSNE with classtpx coloring



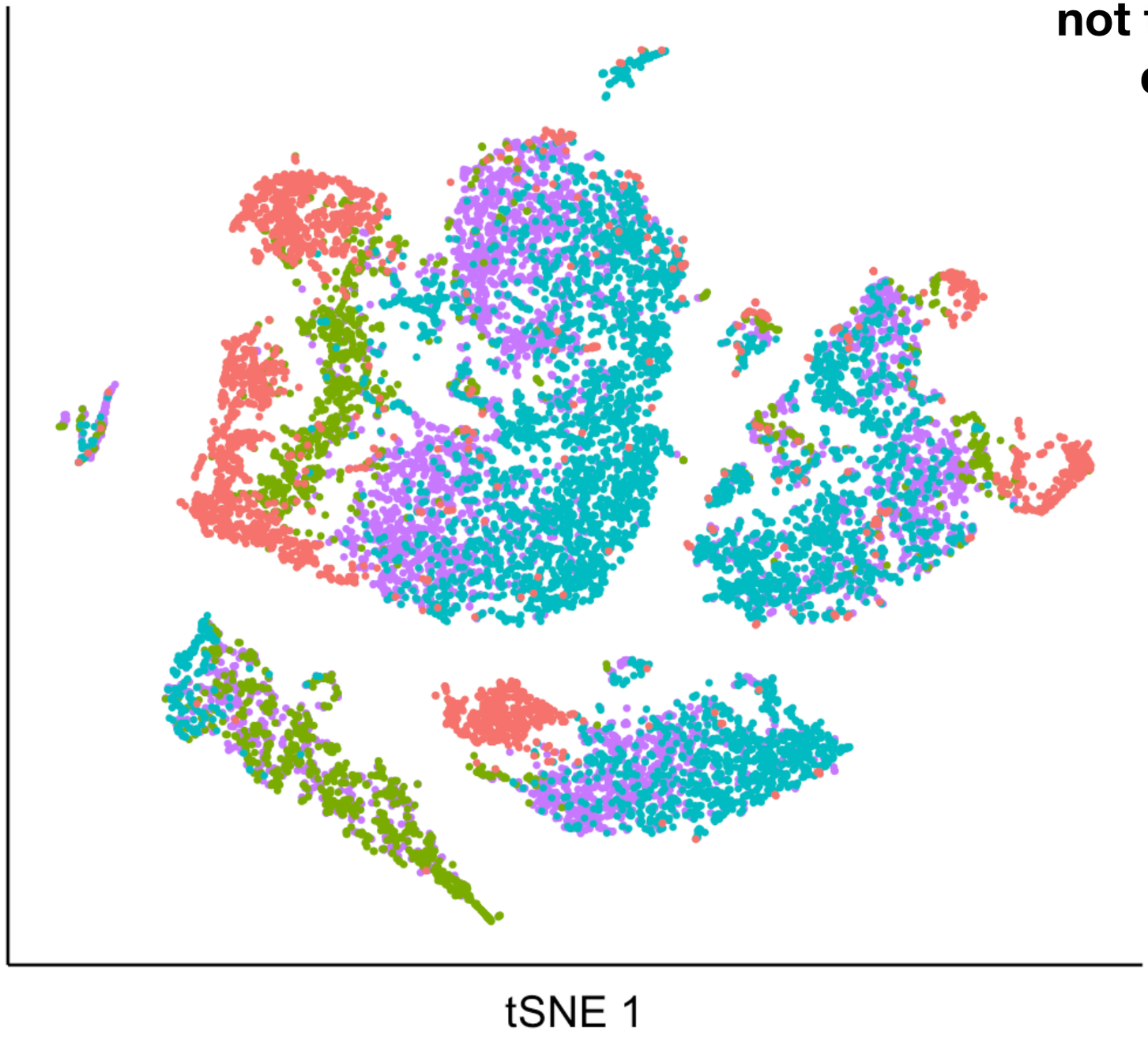
Merkel Cell Carcinoma Study

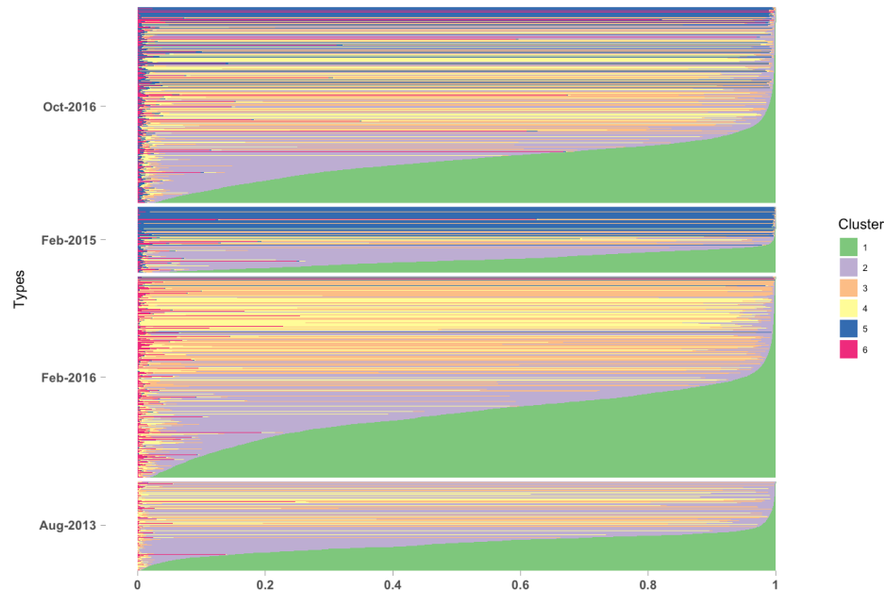


	Aug 2013	Feb 2015	Feb 2016	Oct 2016
Time point:	**Pre-ImmunoTx**	**Failing T cells & α-PD-1**	**Responding to T cells & α-PD-1 & α-CTLA**	**Failing T cells & α-PD-1 & α-CTLA**
Samples for 10X-Genomics	Single-cell tumor digest (viable) TIL (viable) PBMC	PBMC	PBMC	Single-cell tumor digest (viable) TIL (viable) PBMC
Complementary Samples	BX: IHC. TIL, PBMC: T cell phenotype, Adaptive	PBMC: T cell phenotype, adaptive	BX: IHC TIL: Adaptive PBMC: T cell phenotype, adaptive	BX: IHC. TIL, PBMC: T cell phenotype, Adaptive



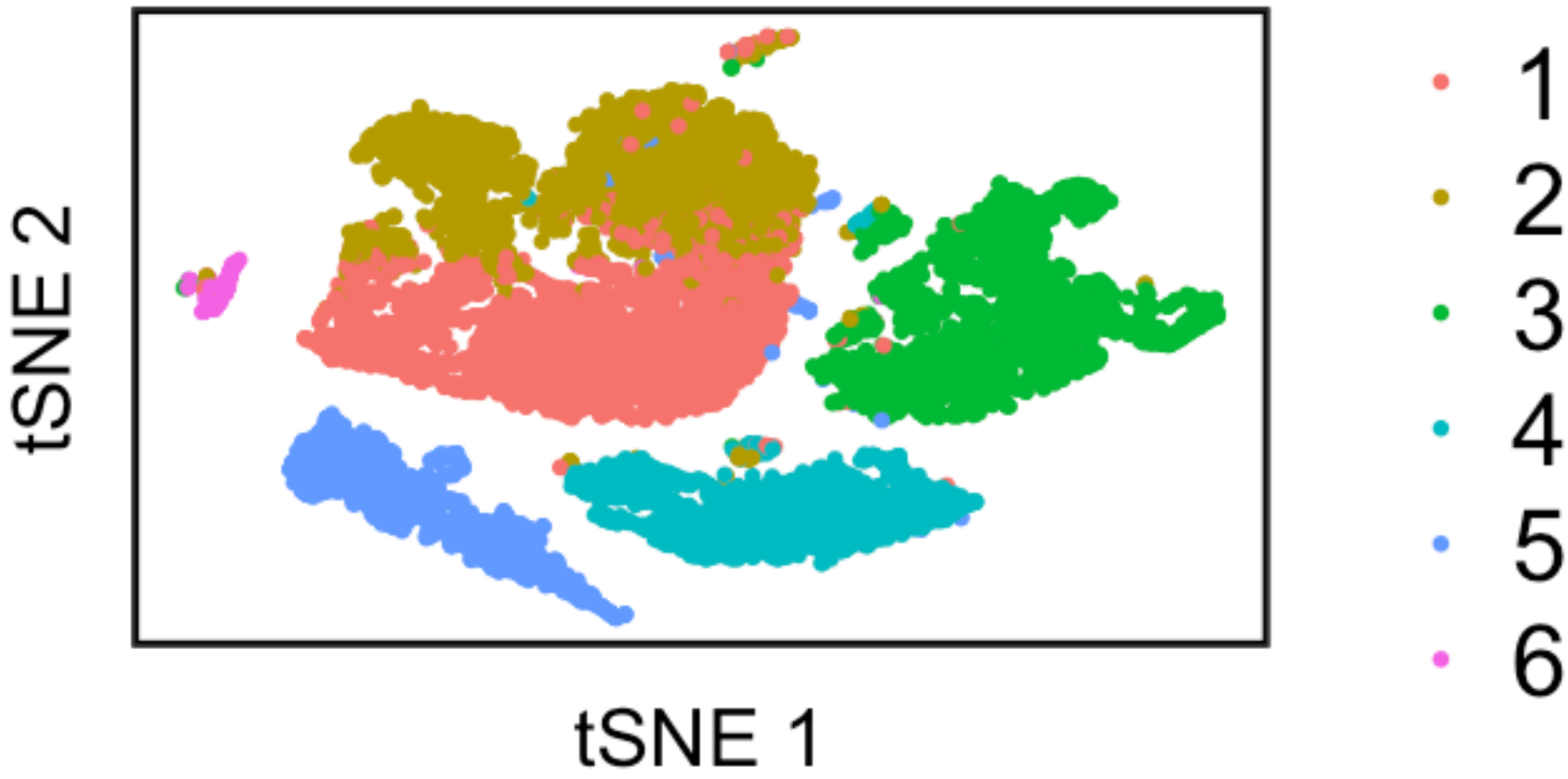
Time batch effect: not the main cluster driving force



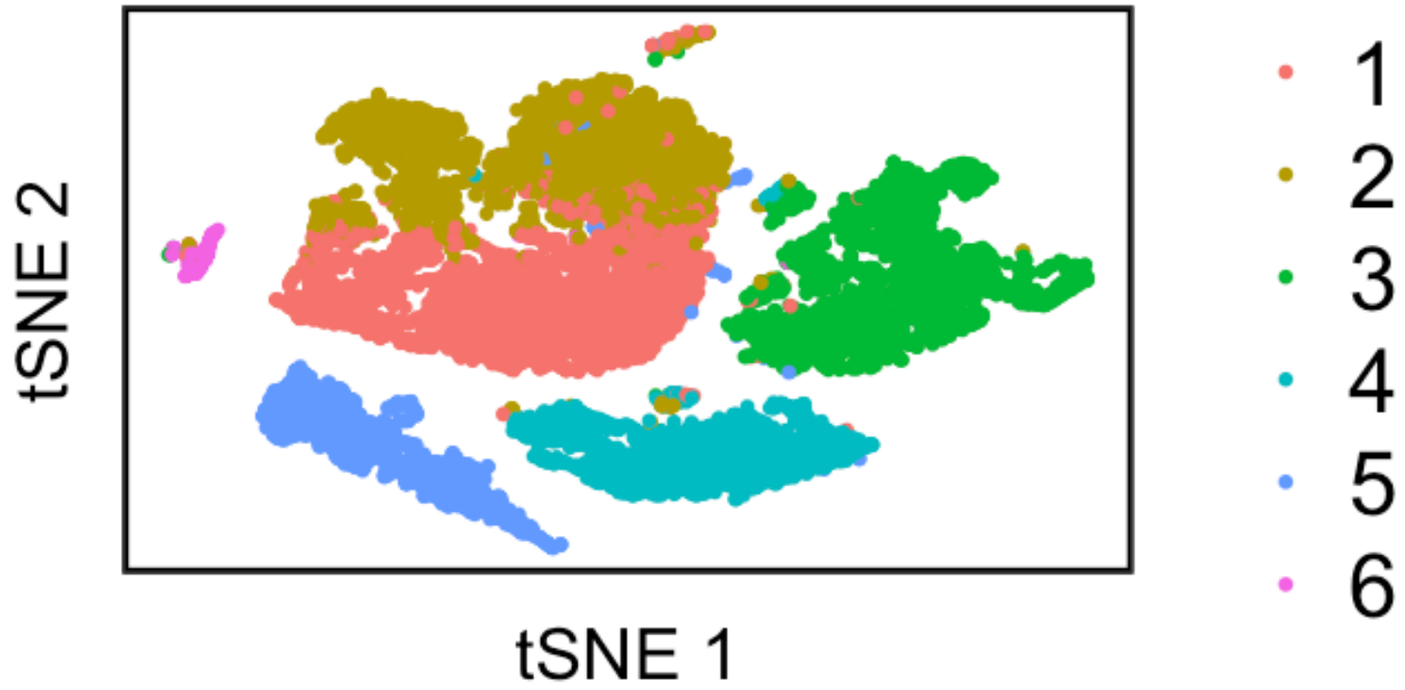


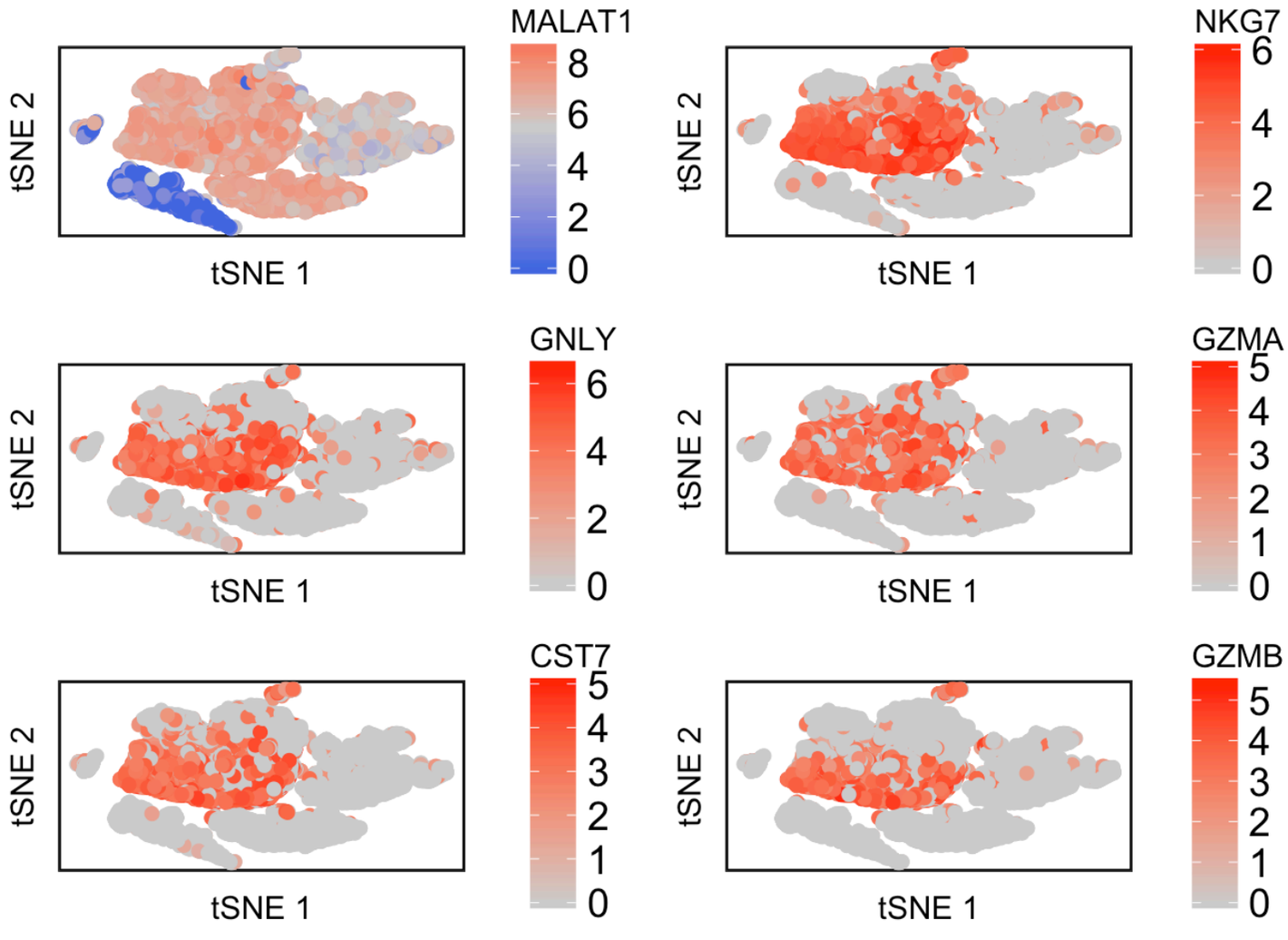
Learning clusters using unsupervised CountClust

Mapping our model clusters on t-SNE plot of the cells



Cluster 1

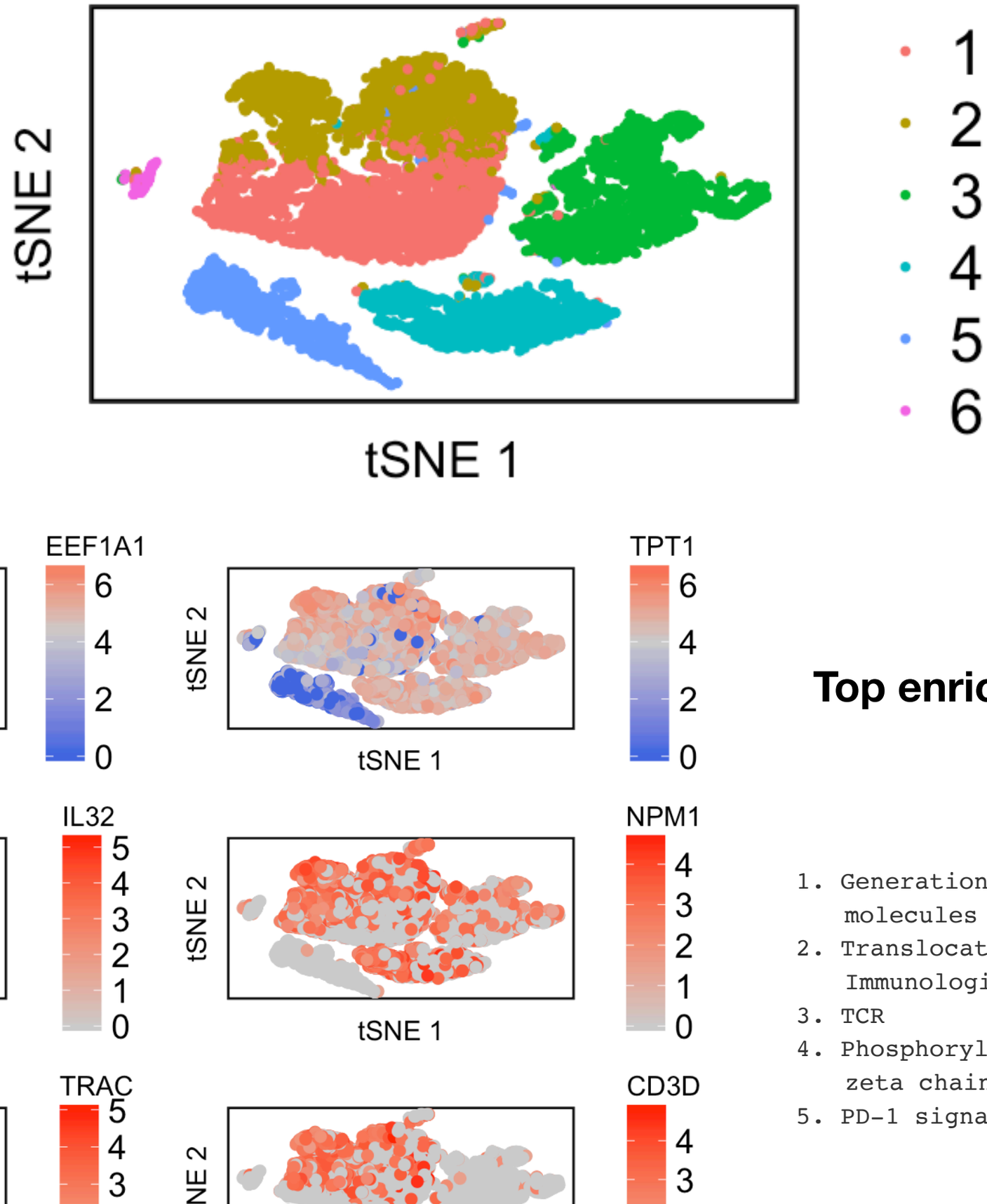




Top enriched pathways

1. Natural killer cell mediated cytotoxicity - Homo sapiens (human)
2. Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell
3. Immune System
4. IL12-mediated signaling events
5. Downstream signaling in naïve CD8+ T cells

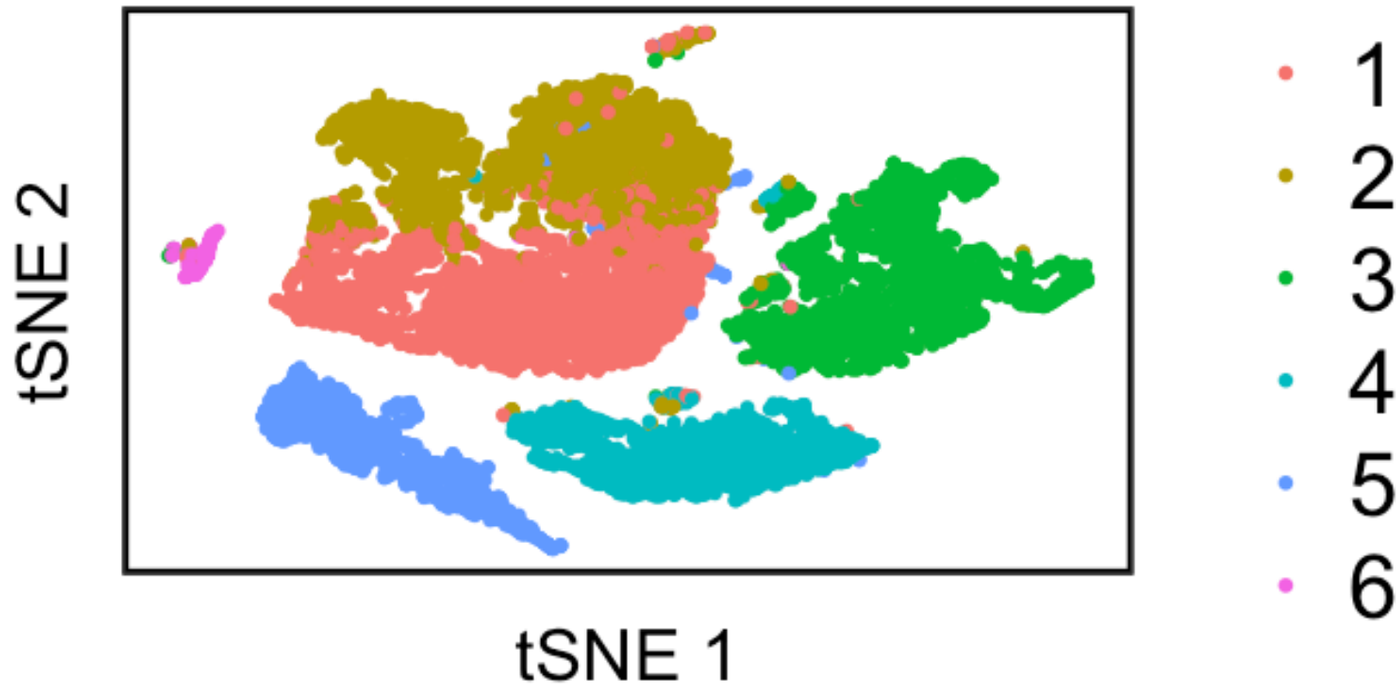
Cluster 2

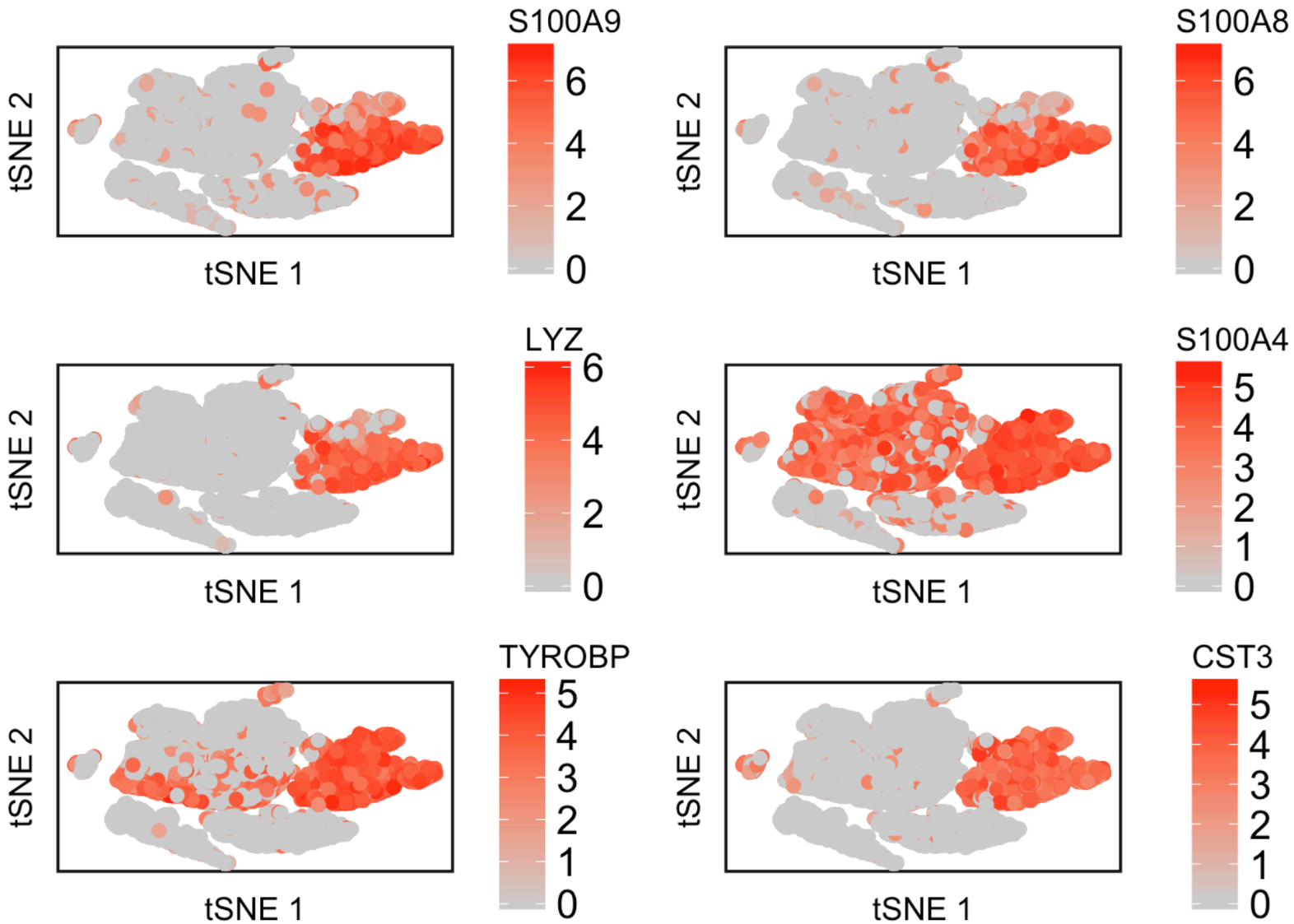


Top enriched pathways

1. Generation of second messenger molecules
2. Translocation of ZAP-70 to Immunological synapse
3. TCR
4. Phosphorylation of CD3 and TCR zeta chains
5. PD-1 signaling

Cluster 3

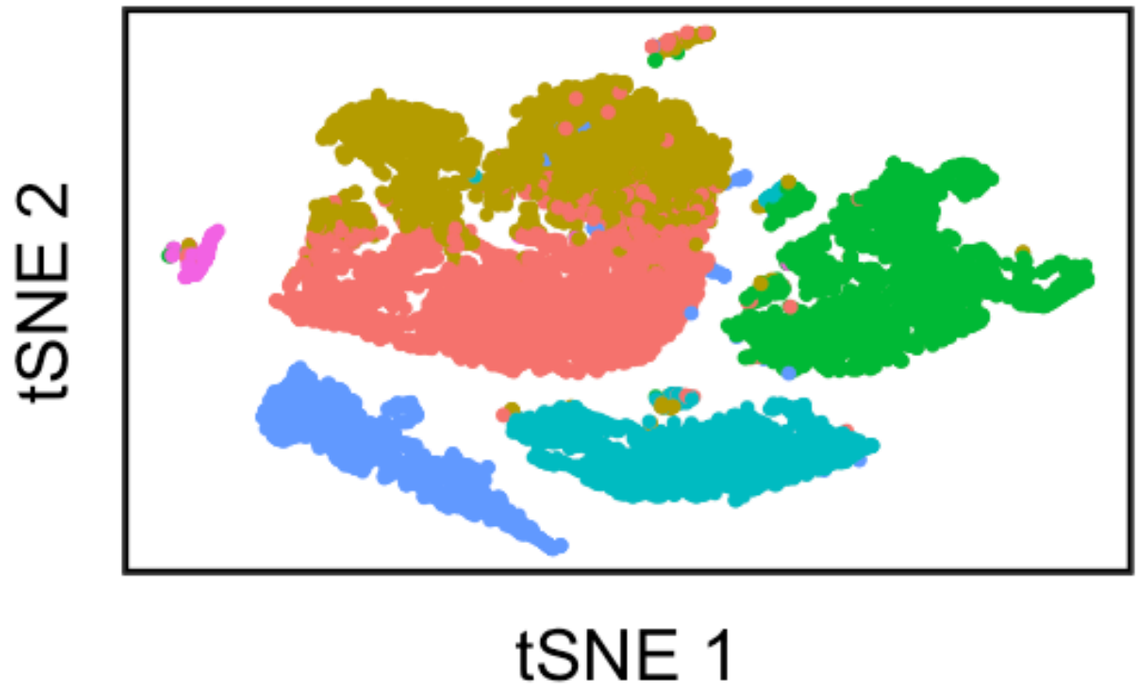


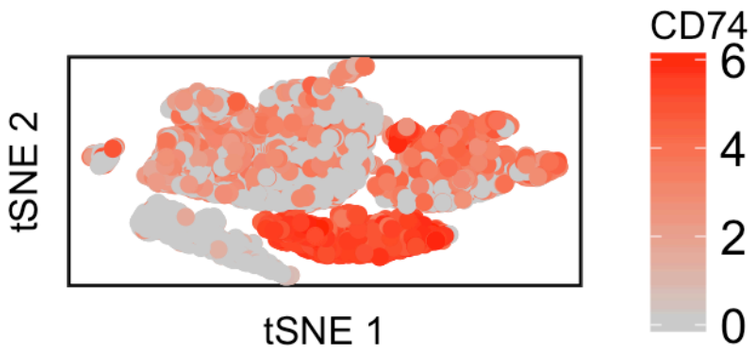


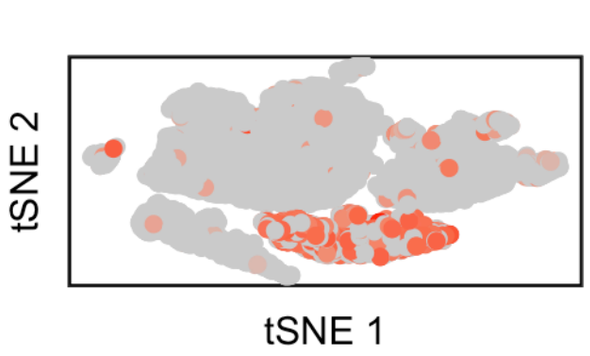
Top enriched pathways

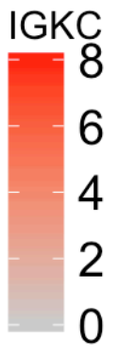
1. Neutrophil degranulation
2. Innate Immune System
3. Immune System
4. Toll-Like Receptors Cascades
5. Endogenous TLR signaling
6. Regulation of TLR by endogenous ligand

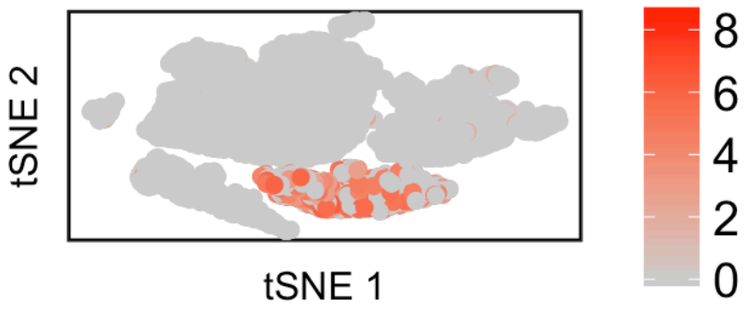
Cluster 4

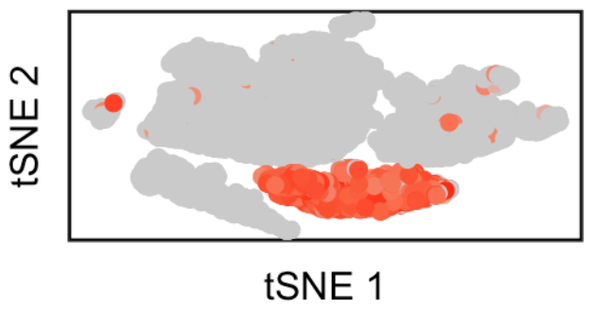


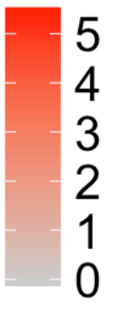


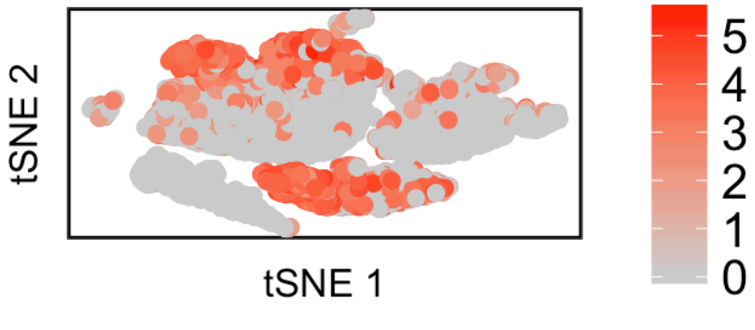


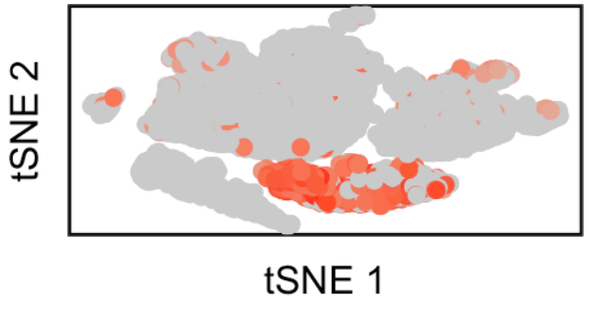


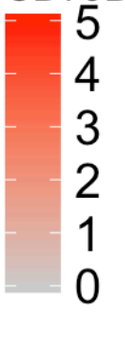








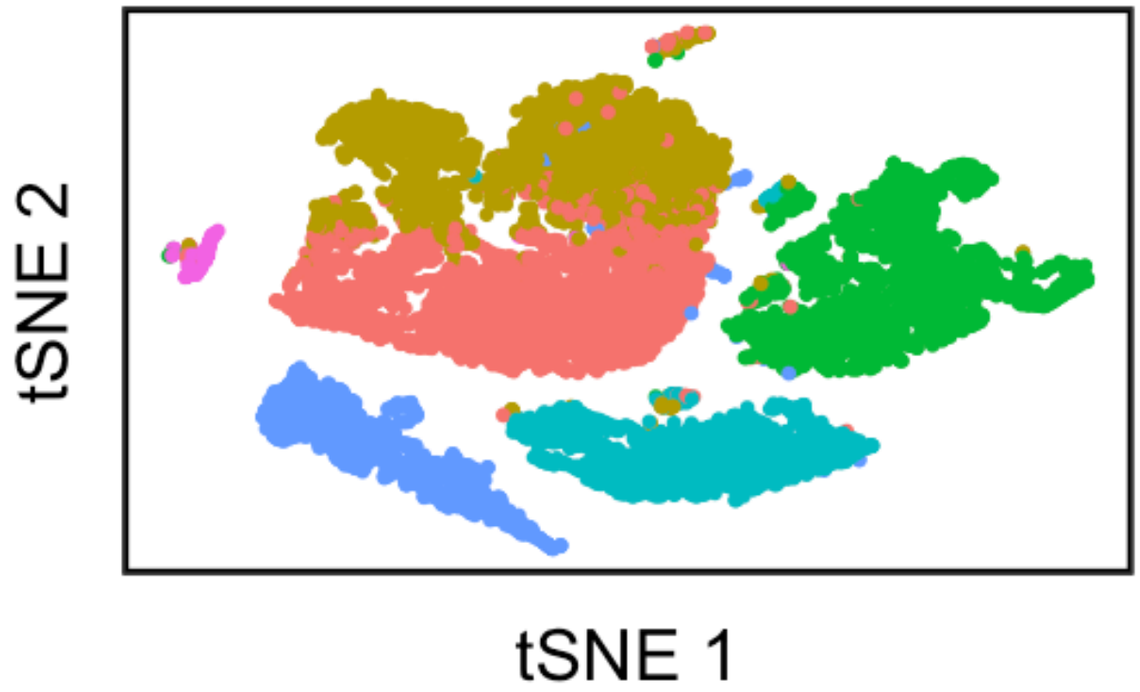


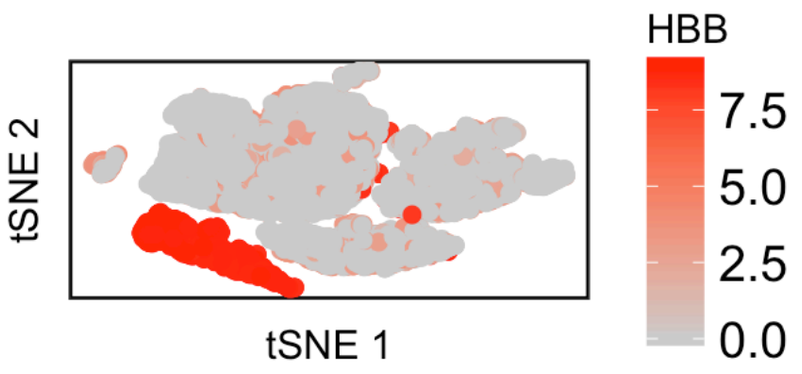


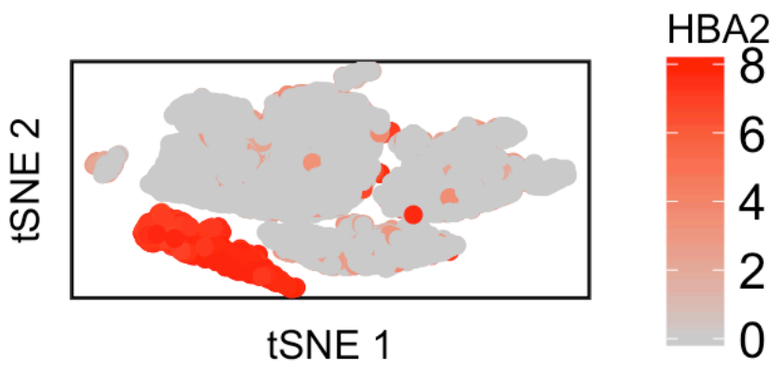
Top enriched pathways

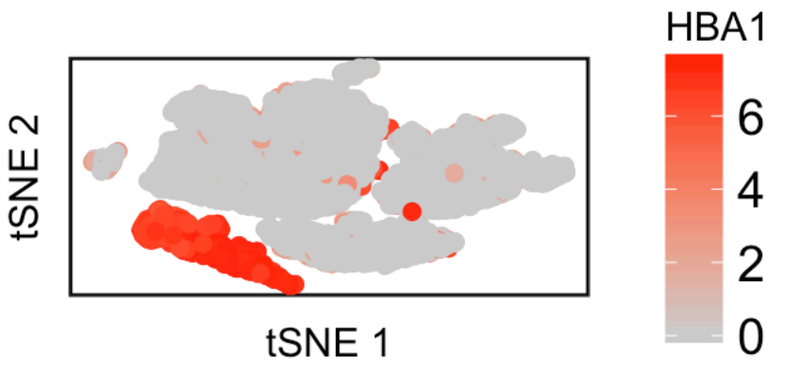
1. CD22 mediated BCR regulation
2. Antigen activates B Cell Receptor (BCR) leading to generation of second messengers
3. Cell surface interactions at the vascular wall
4. BCR
5. Signaling by the B Cell Receptor (BCR)

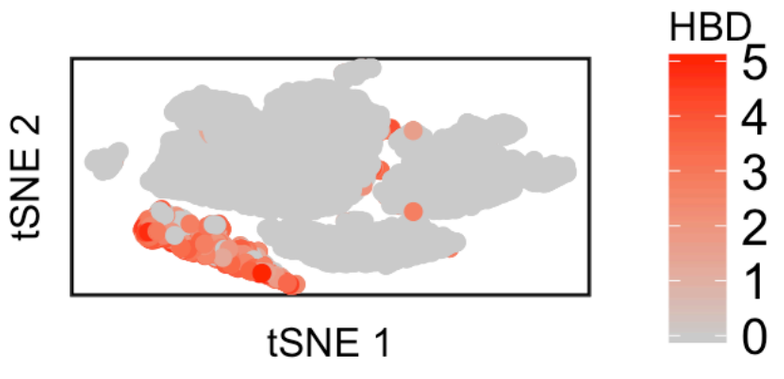
Cluster 5

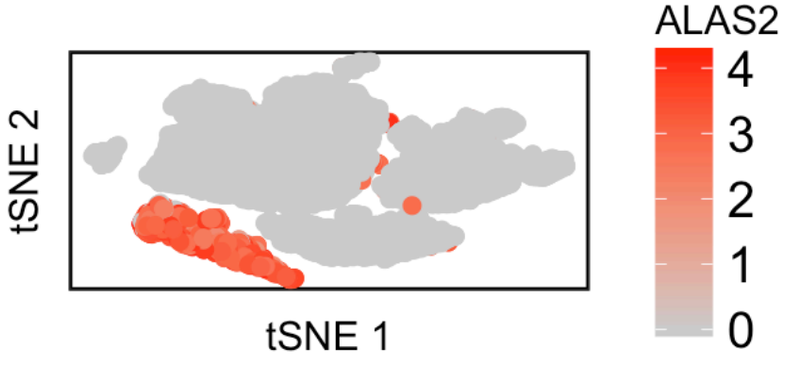


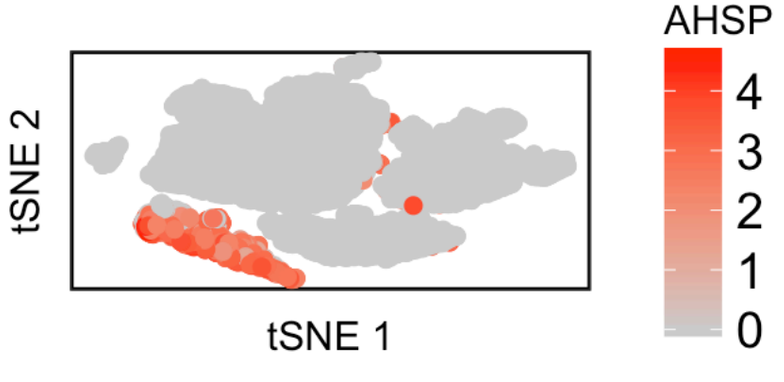








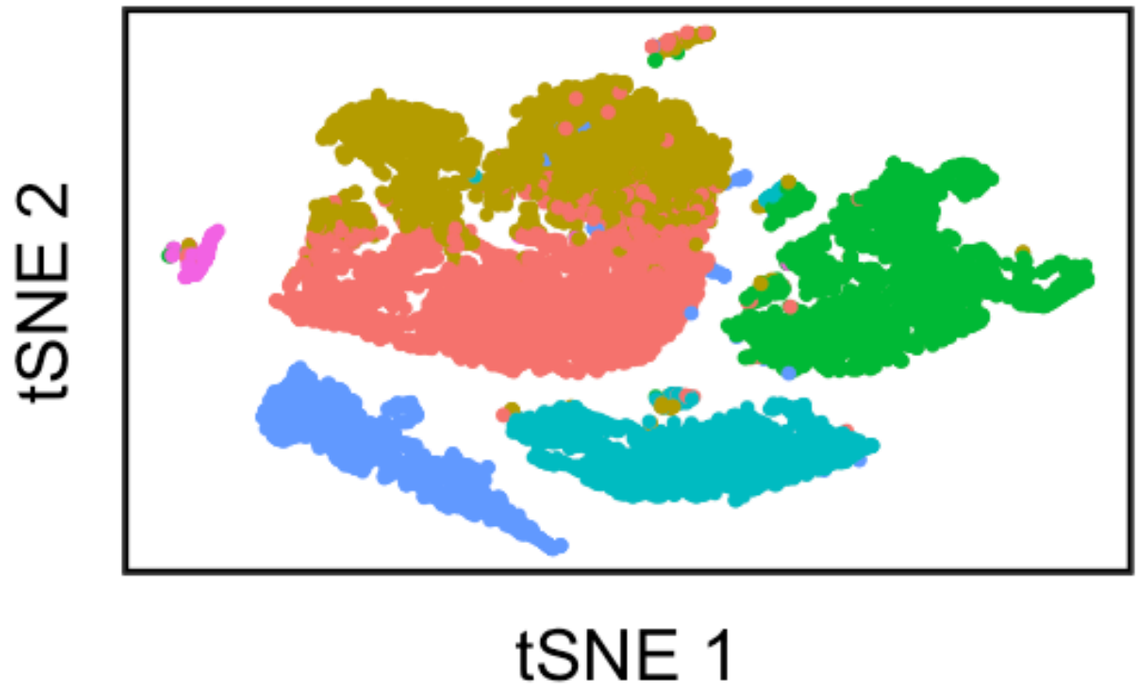


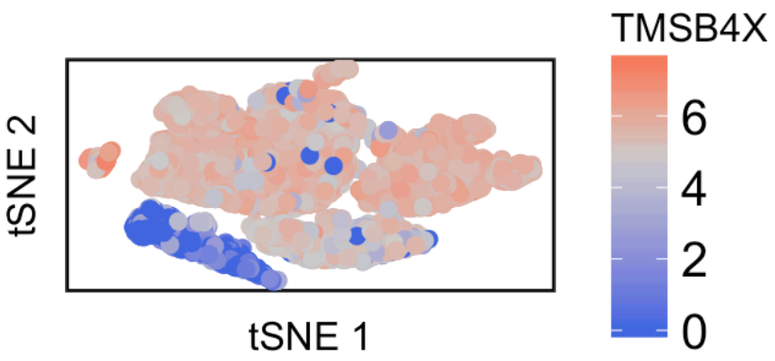


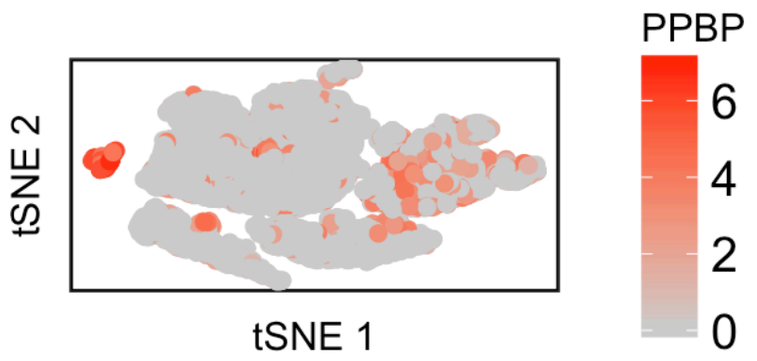
Top enriched pathways

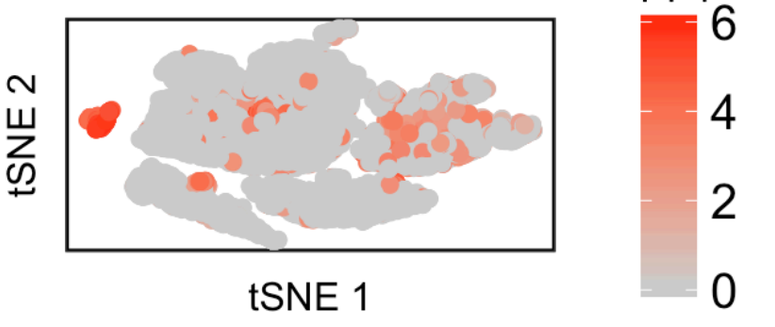
1. hemoglobins chaperone
2. Erythrocytes take up oxygen and release carbon dioxide
3. Erythrocytes take up carbon dioxide and release oxygen
4. O2/CO2 exchange in erythrocytes
5. Malaria - Homo sapiens (human)

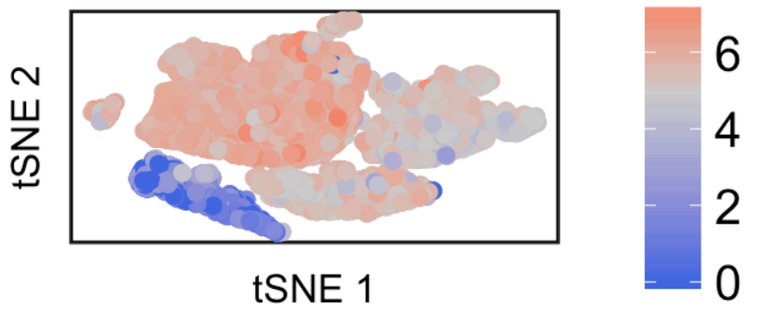
Cluster 6

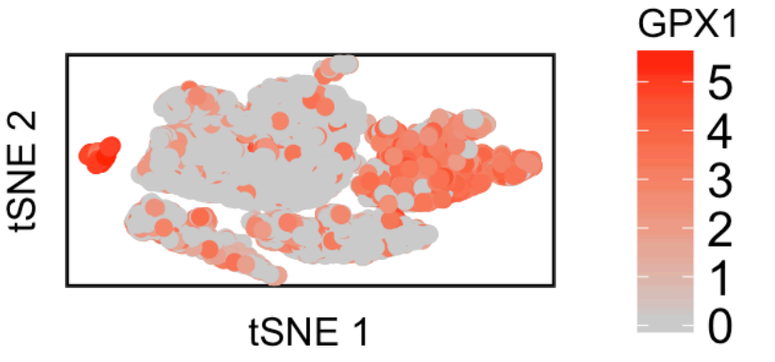


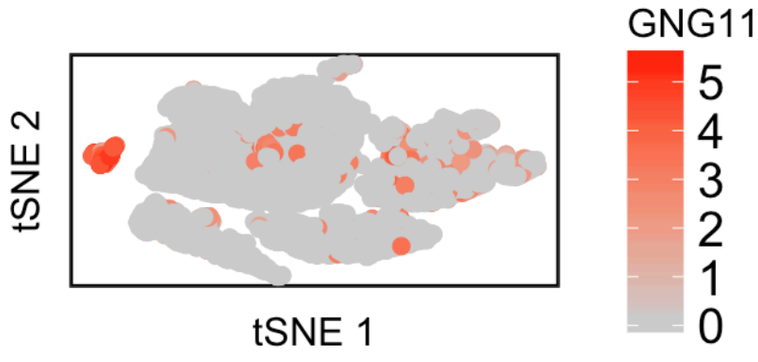










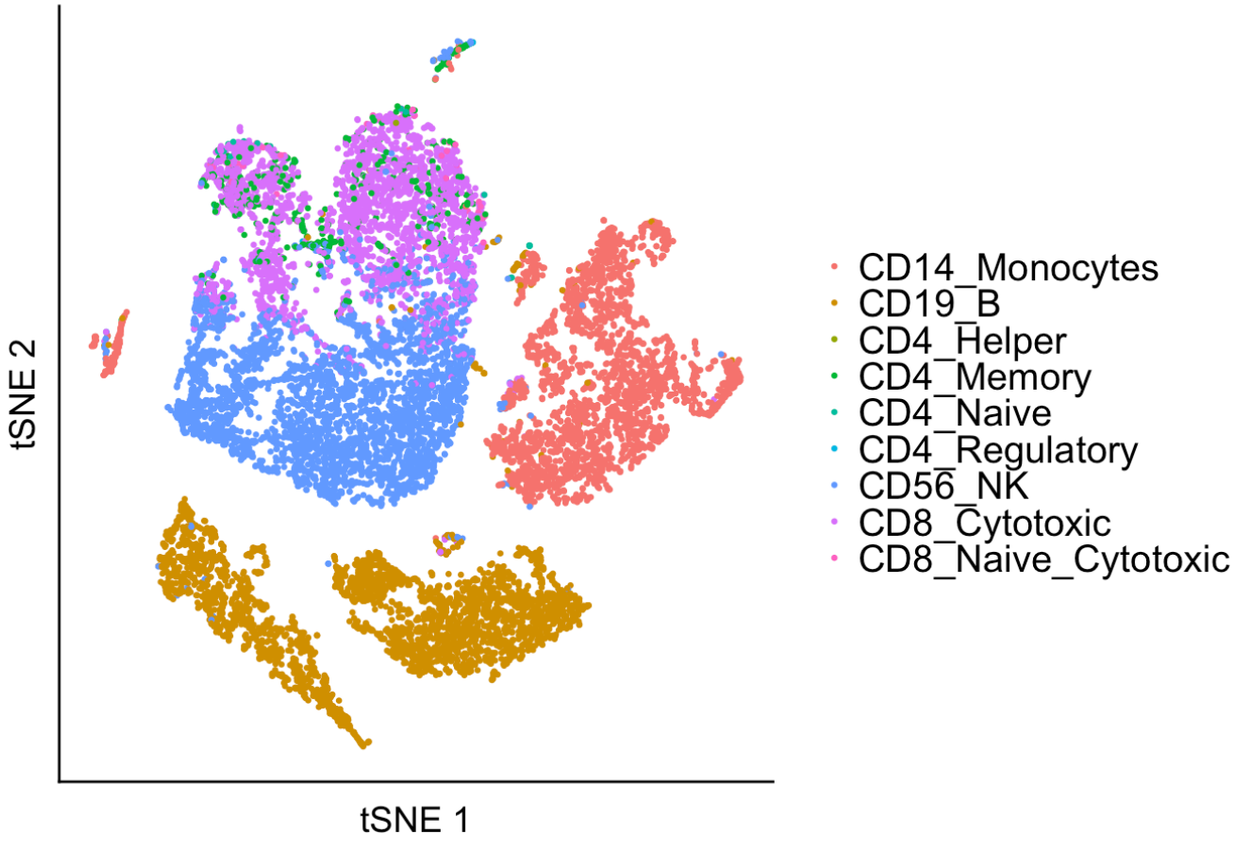


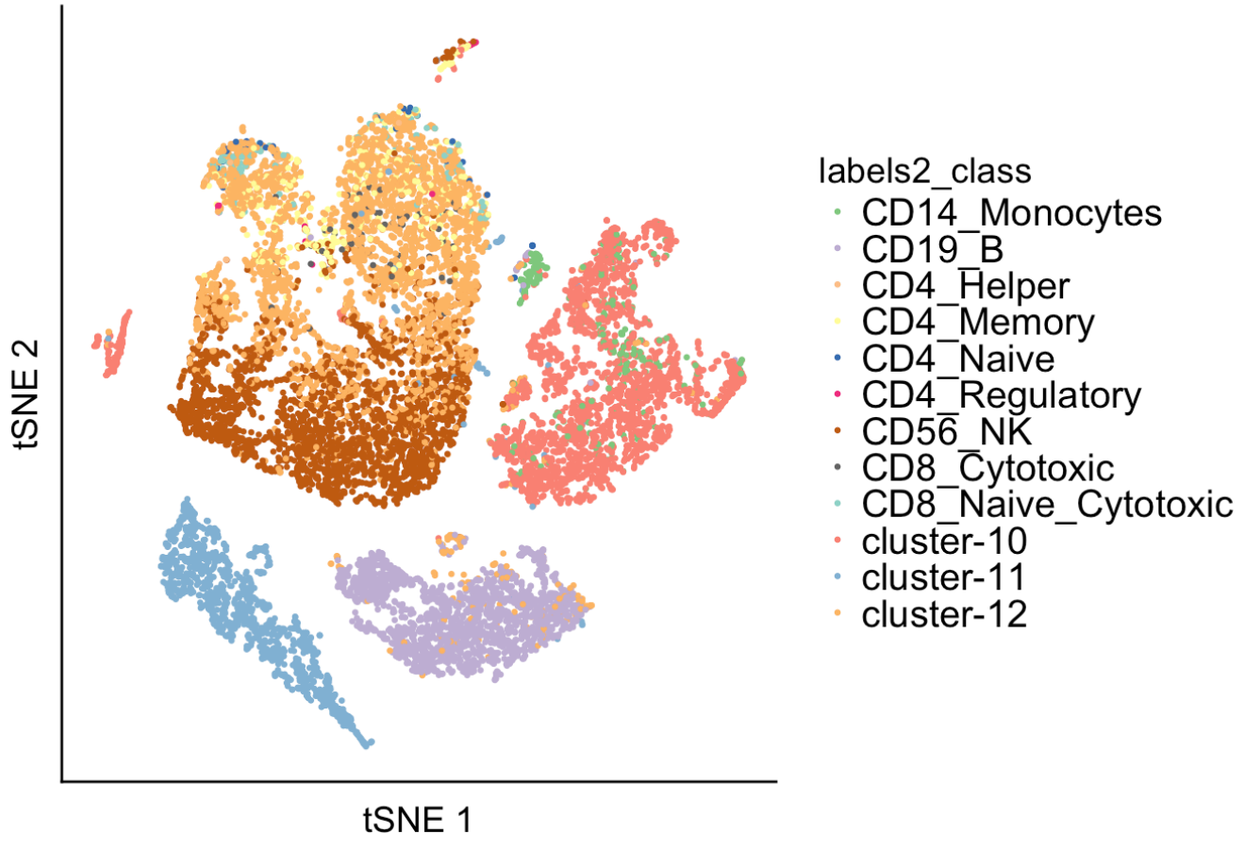
Top enriched pathways

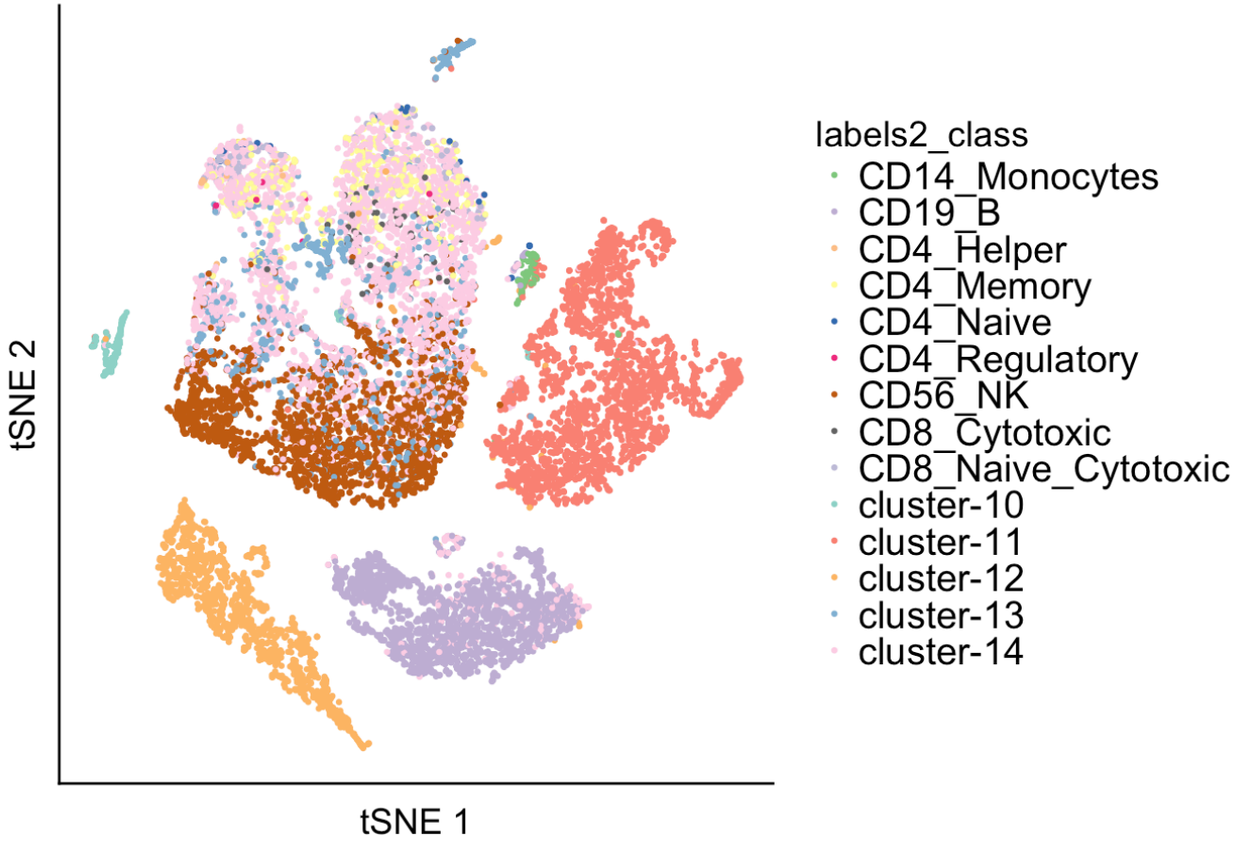
1. Platelet activation, signaling and aggregation
2. Platelet degranulation
3. Response to elevated platelet cytosolic Ca2+
4. Hemostasis
5. RHO GTPases activate PKNs

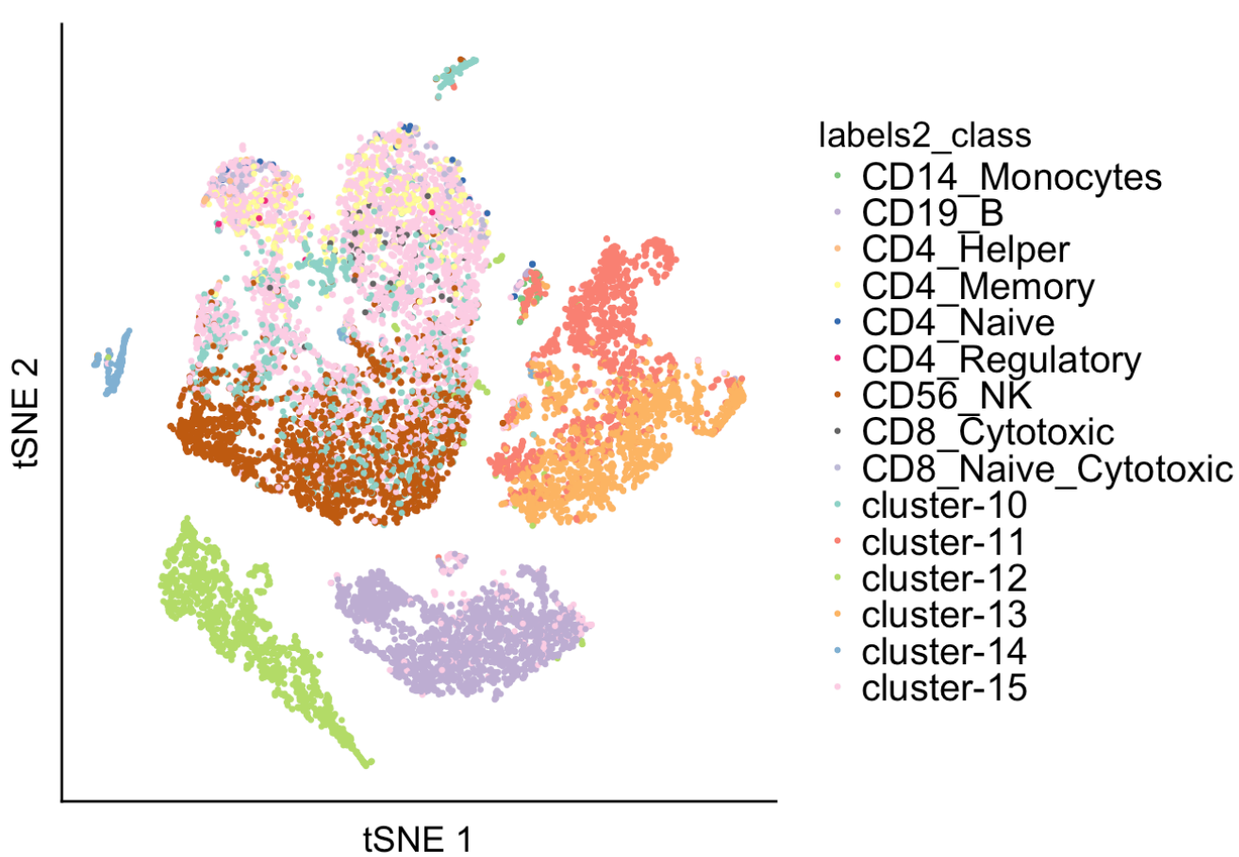
Learning cellular types using partially supervised CountClust

MCC data (with Zheng et al sorted immune cell training)


tSNE 2
tSNE 1
CD14_Monocytes
CD19_B
CD4_Helper
CD4_Memory
CD4_Naive
CD4_Regulatory
CD56_NK
CD8_Cytotoxic
CD8_Naive_Cytotoxic

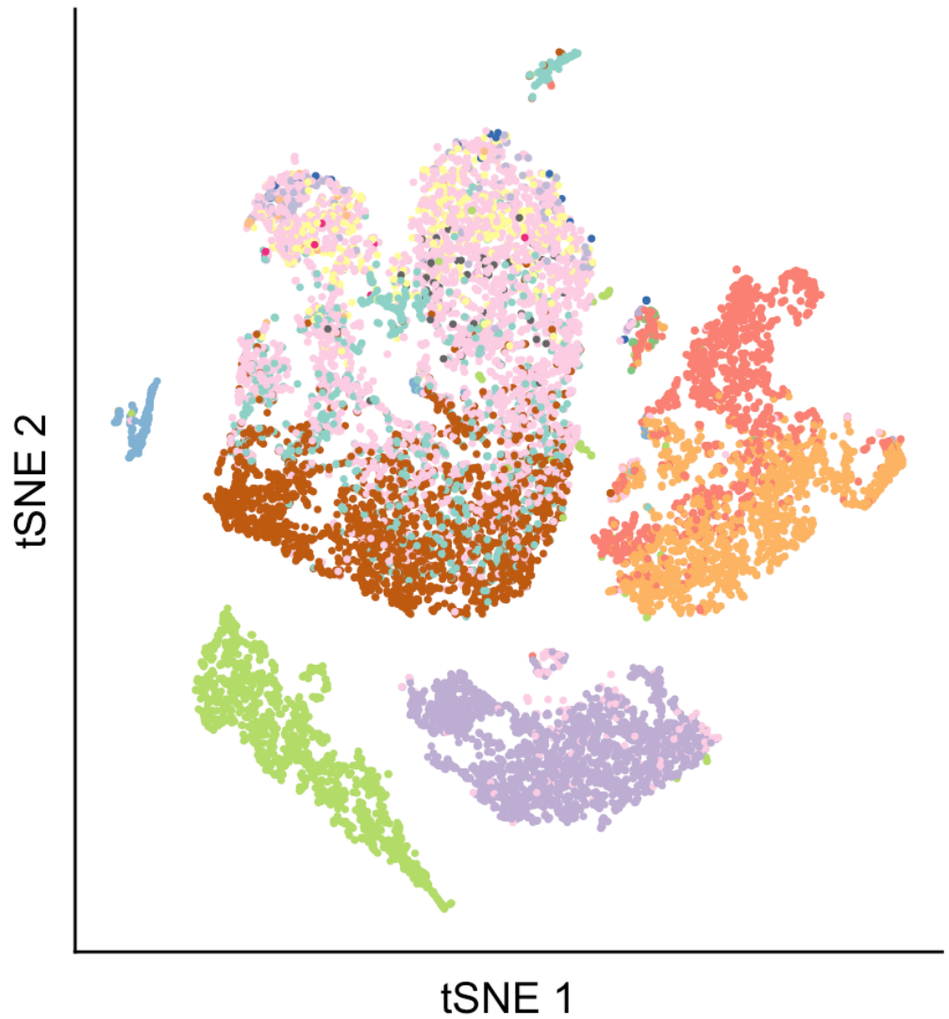

tSNE 2
tSNE 1
labels2_class
CD14_Monocytes
CD19_B
CD4_Helper
CD4_Memory
CD4_Naive
CD4_Regulatory
CD56_NK
CD8_Cytotoxic
CD8_Naive_Cytotoxic
cluster-10
cluster-11
cluster-12


tSNE 2
tSNE 1
labels2_class
CD14_Monocytes
CD19_B
CD4_Helper
CD4_Memory
CD4_Naive
CD4_Regulatory
CD56_NK
CD8_Cytotoxic
CD8_Naive_Cytotoxic
cluster-10
cluster-11
cluster-12
cluster-13
cluster-14

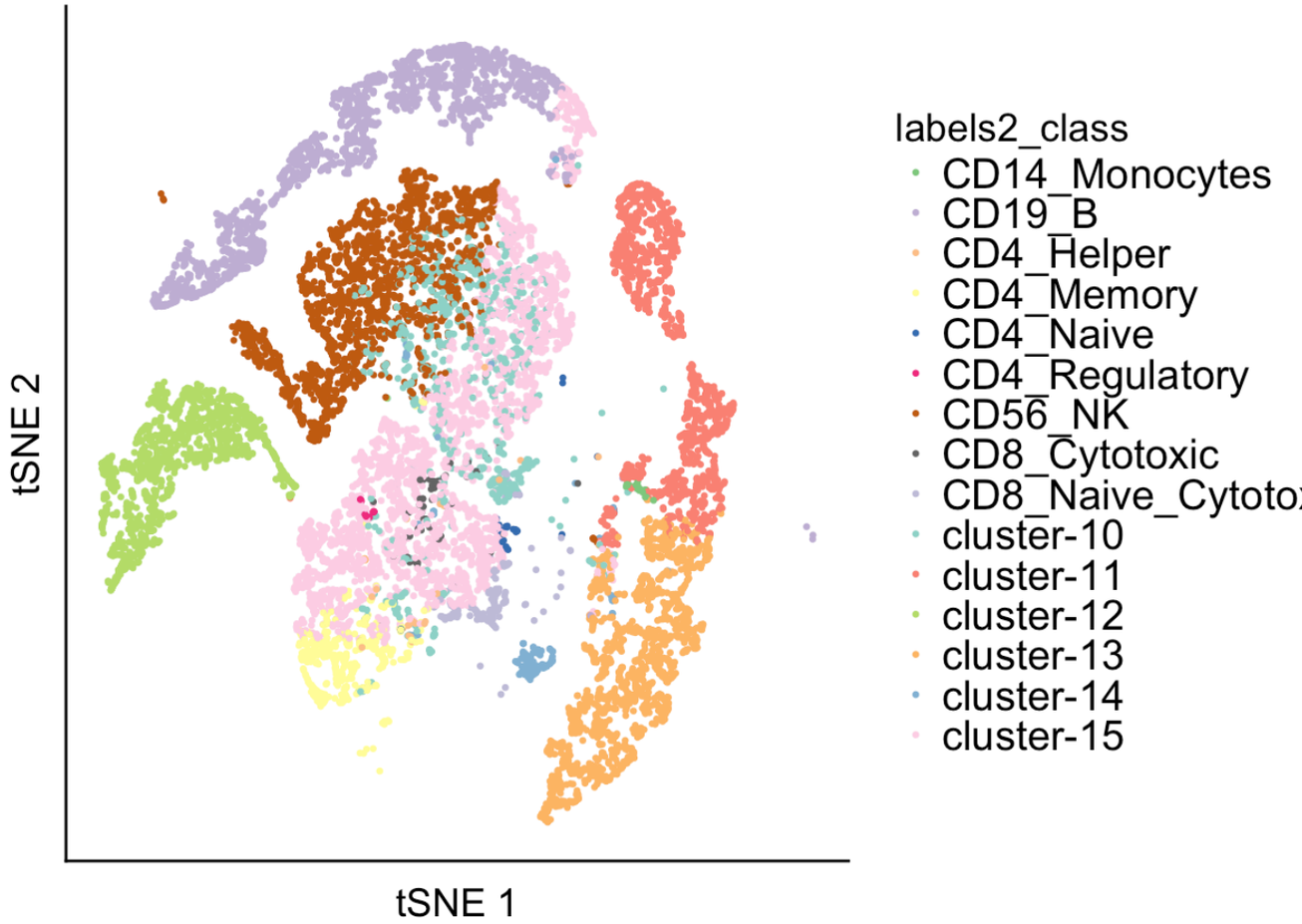

tSNE 2
tSNE 1
labels2_class
CD14_Monocytes
CD19_B
CD4_Helper
CD4_Memory
CD4_Naive
CD4_Regulatory
CD56_NK
CD8_Cytotoxic
CD8_Naive_Cytotoxic
cluster-10
cluster-11
cluster-12
cluster-13
cluster-14
cluster-15

t-SNE before and after supervised CountClust

t-SNE on raw data



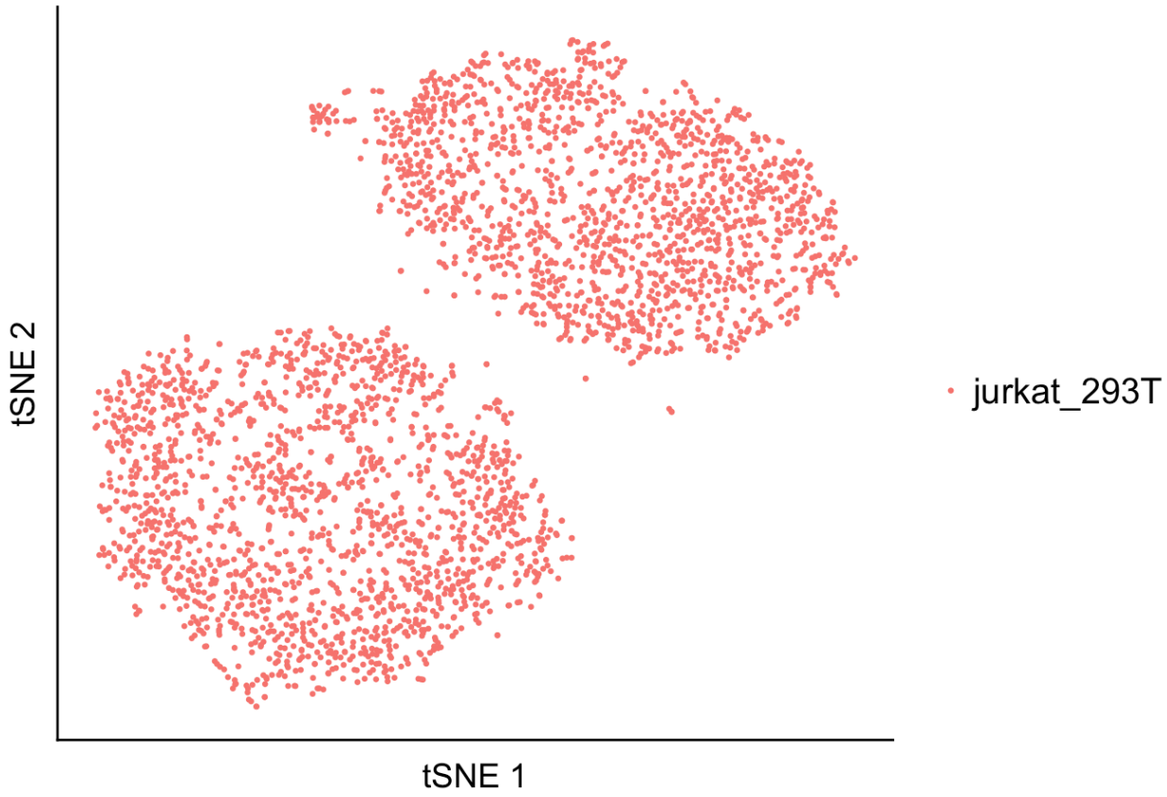
t-SNE on supervised CountClust grades of membership

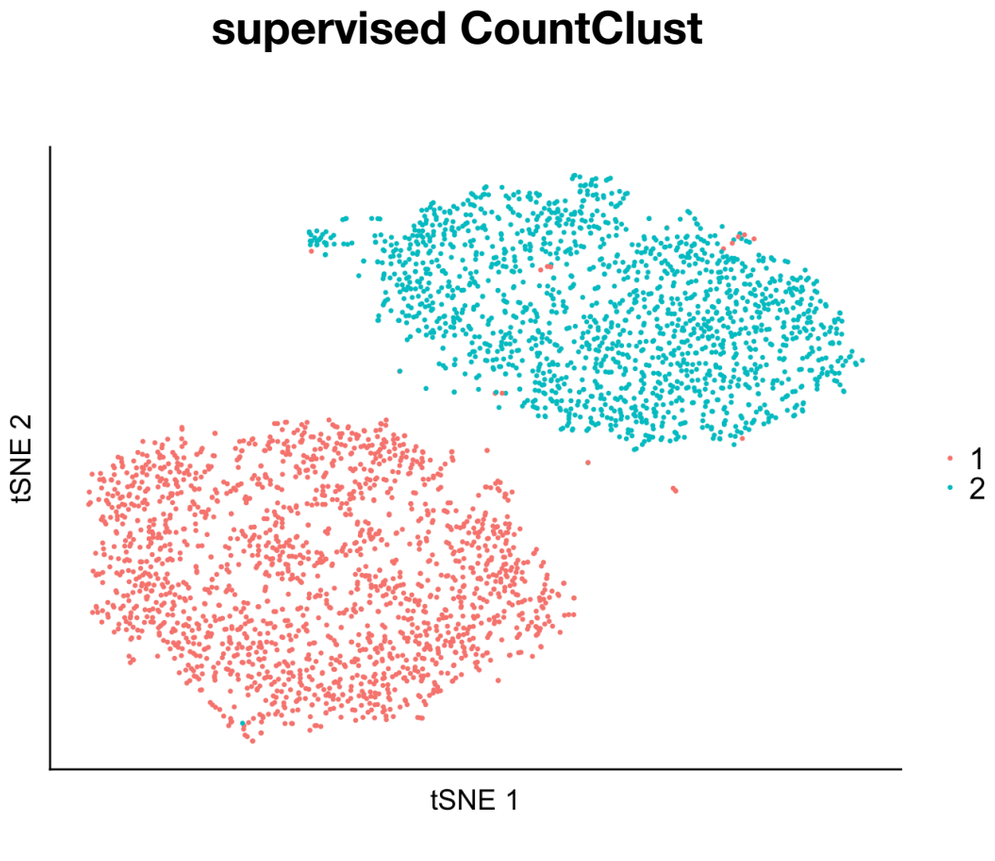


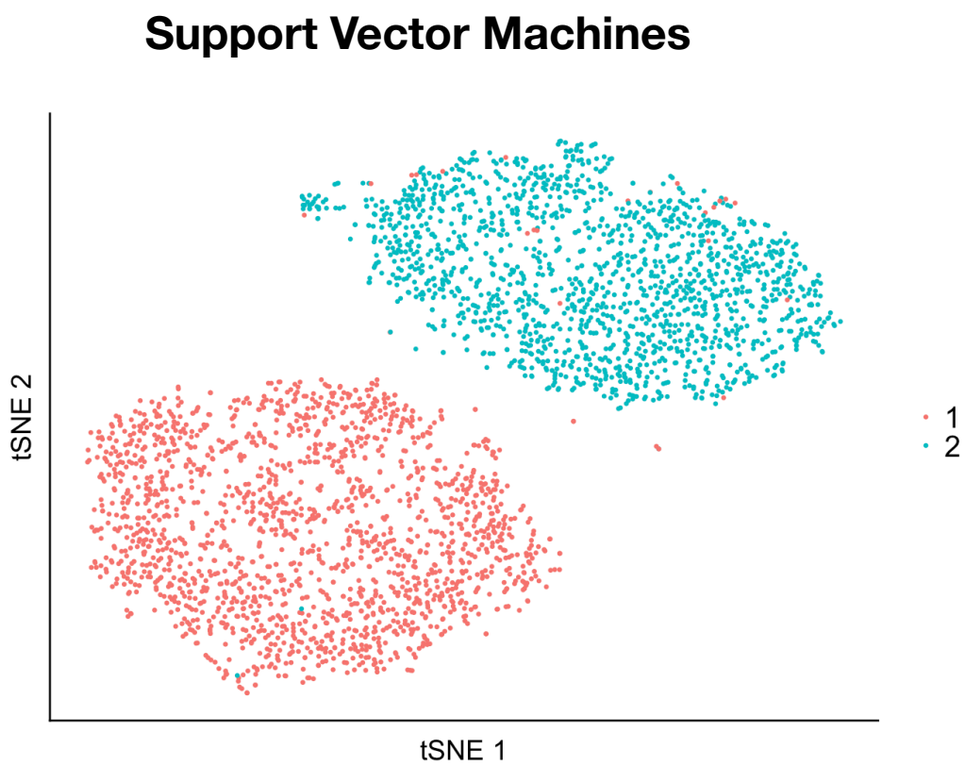
Can Learning be improved by supervised CountClust?

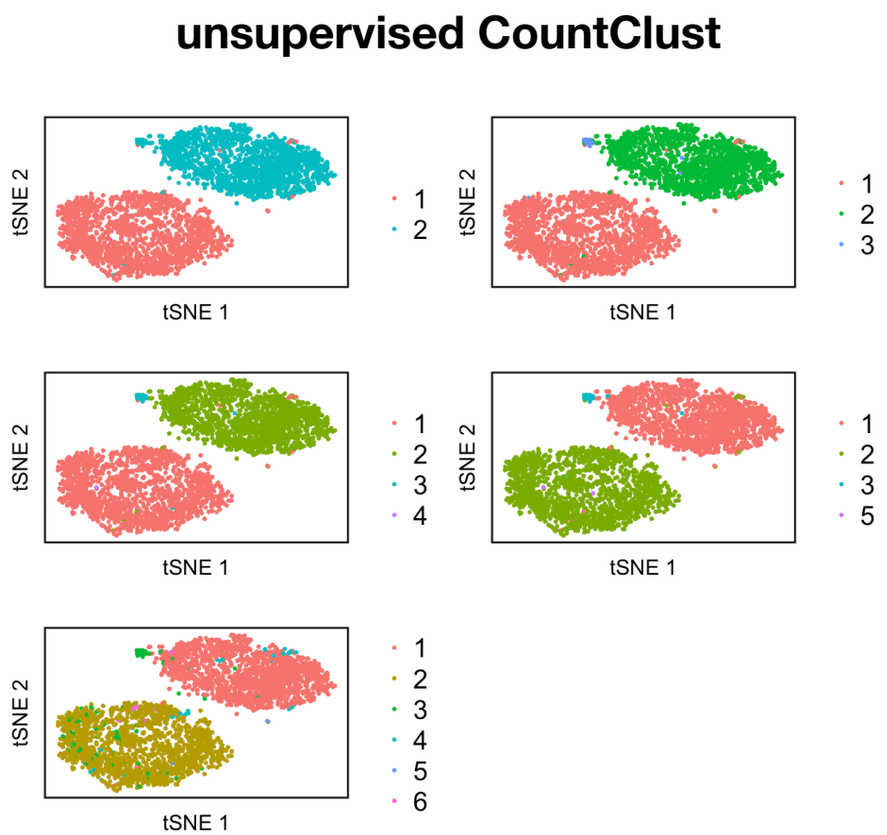
Mixture of cells : Zheng et al 2017

Mixture (50-50) - Jurkat cells + 293T cells



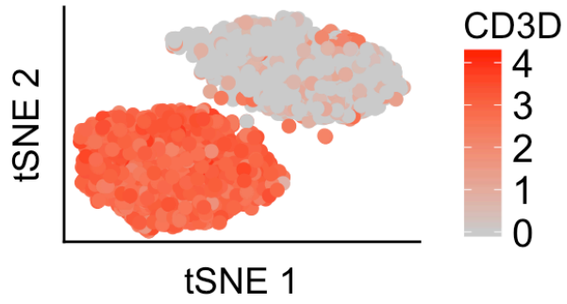


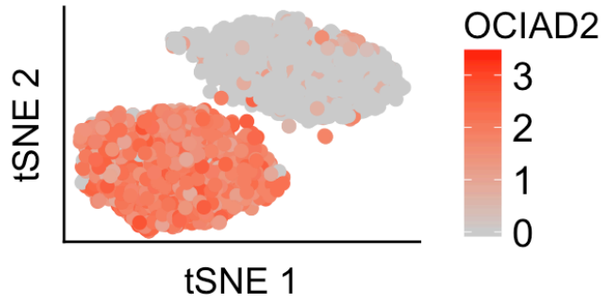


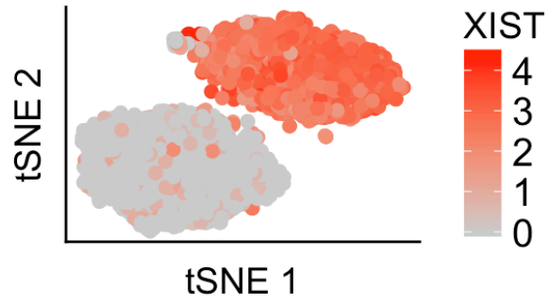


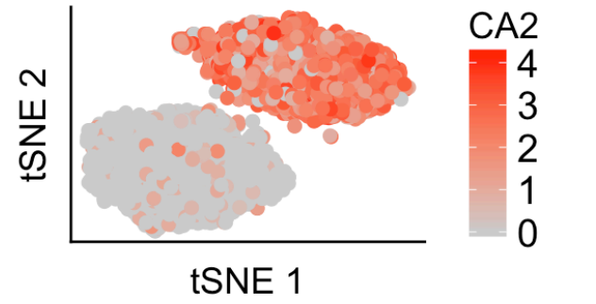
Class 1 - Jurkat cells

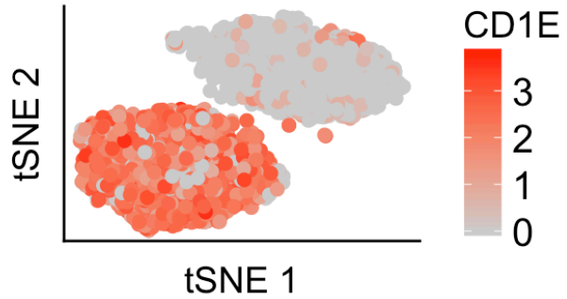
Class 2 - 293T cells

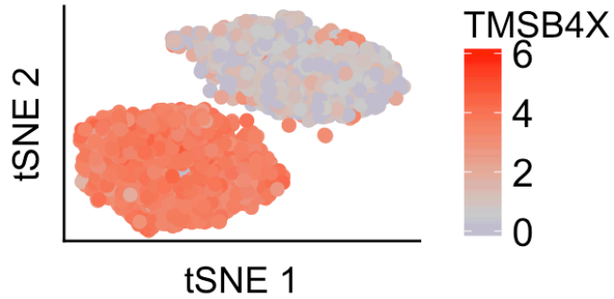


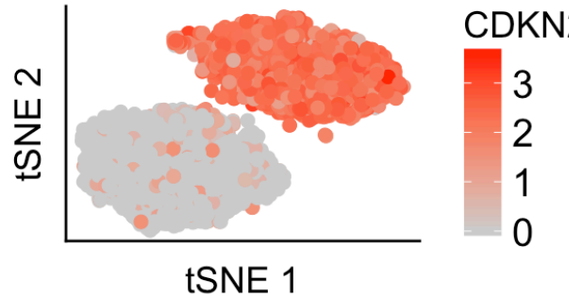


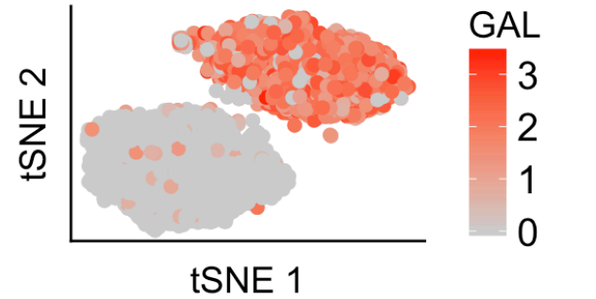


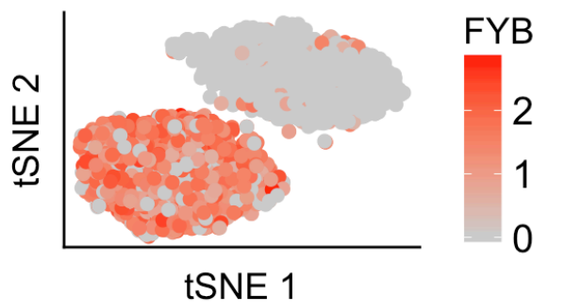


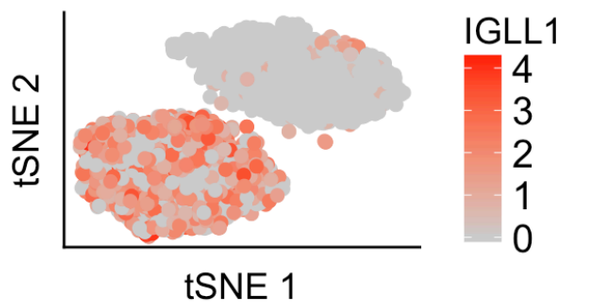


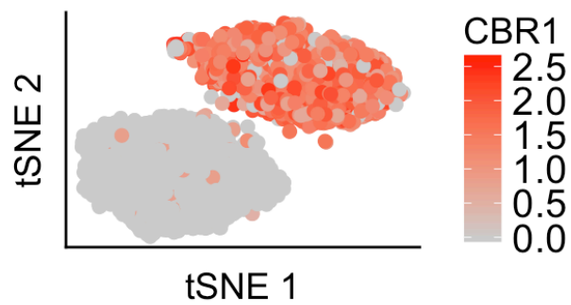


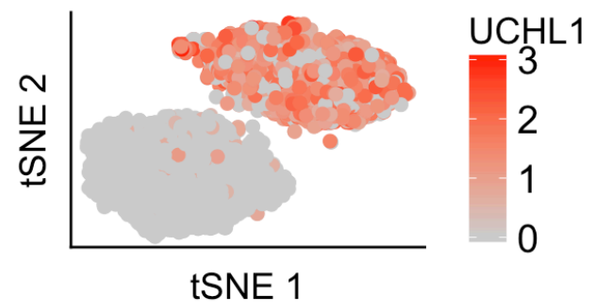




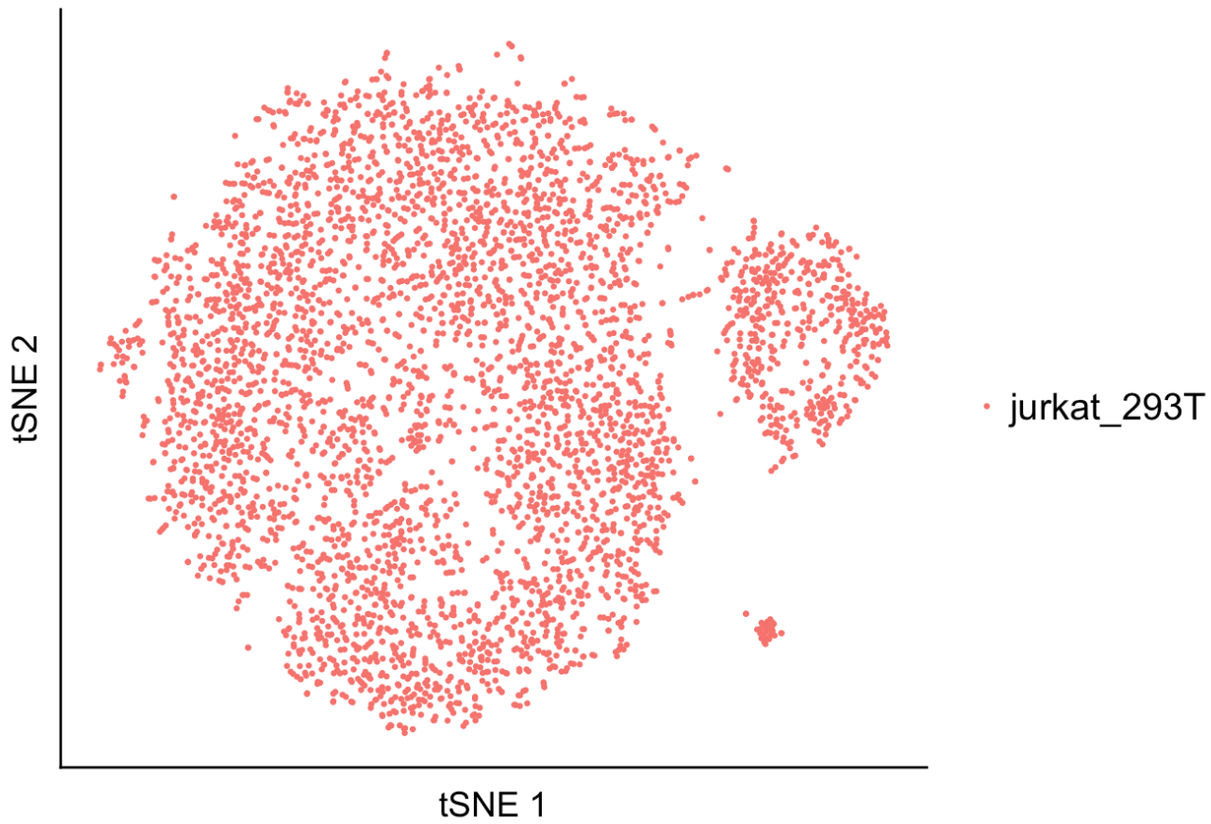


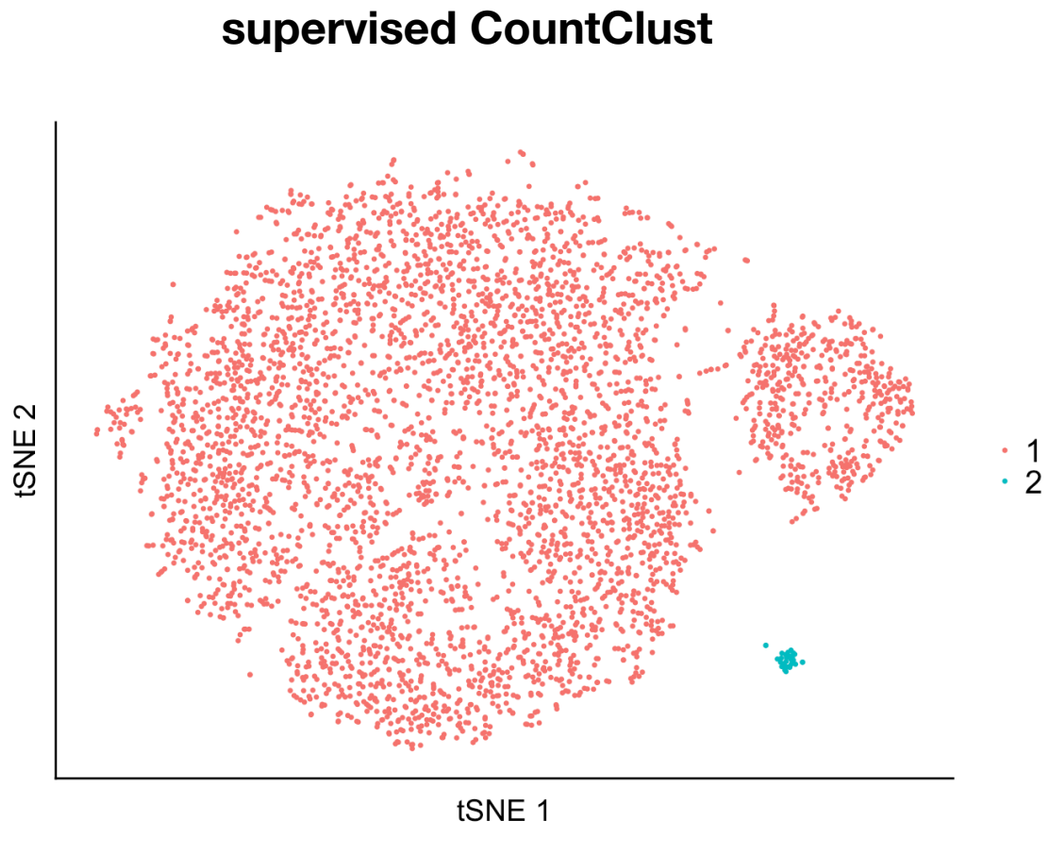


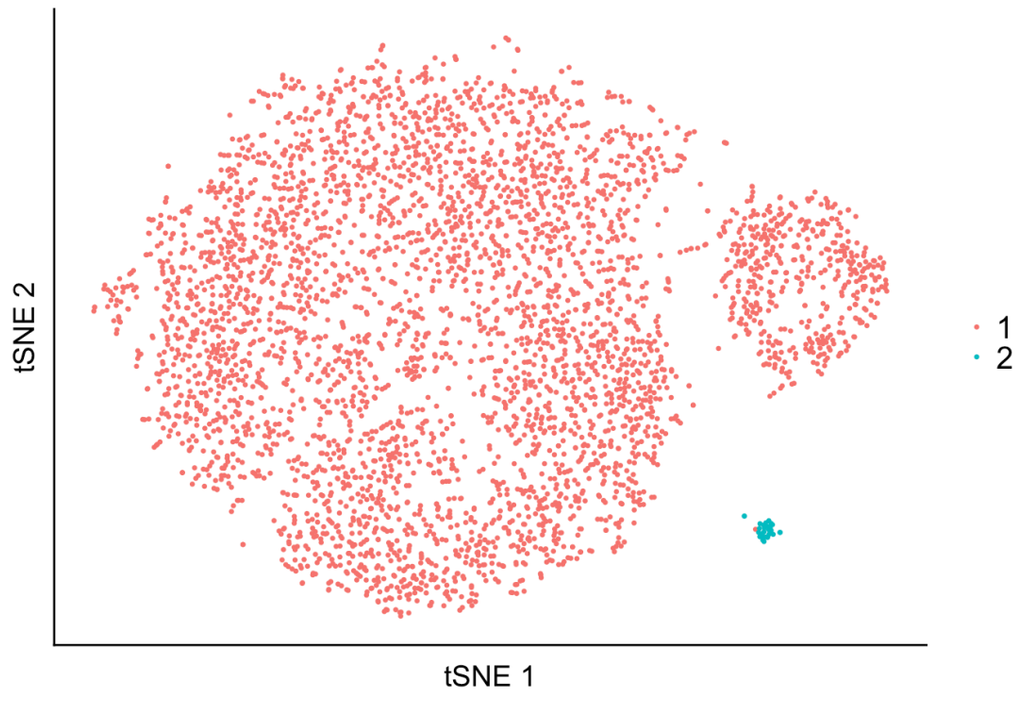


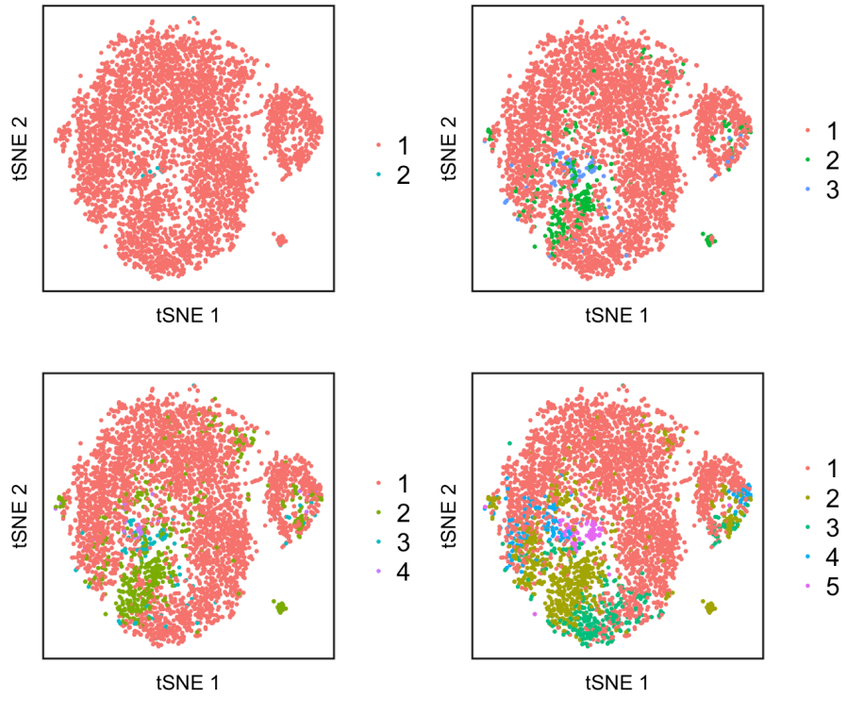


Mixture (99-1) - Jurkat cells + 293T cells

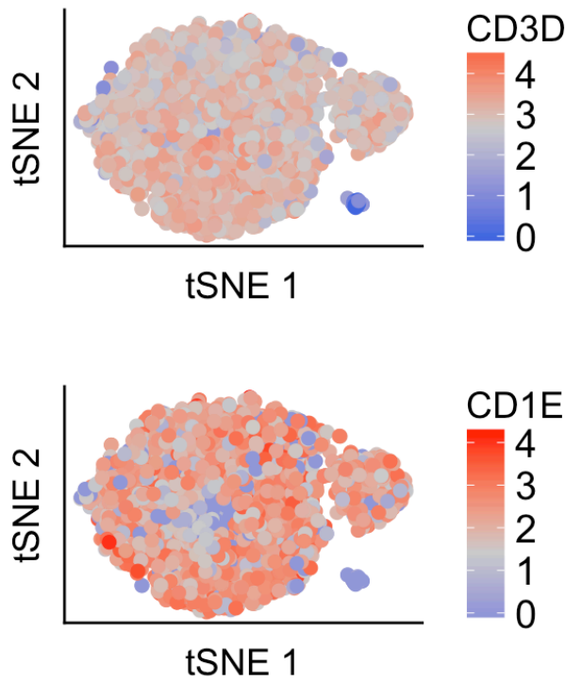


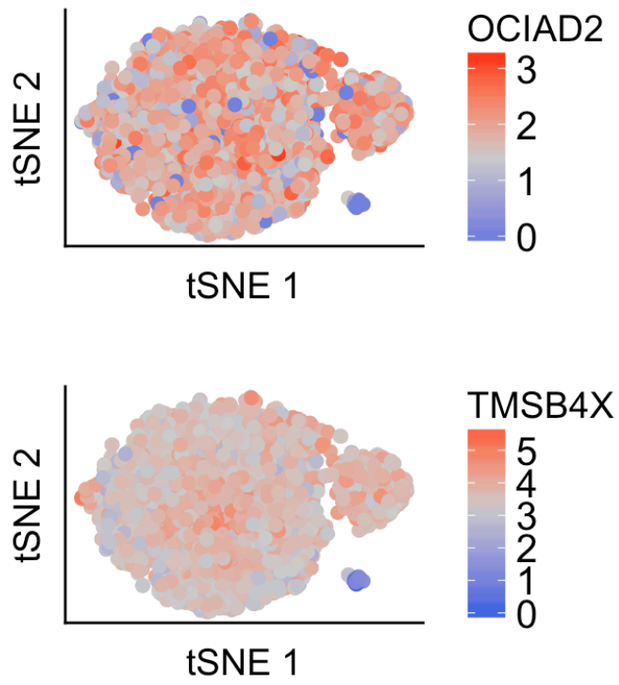


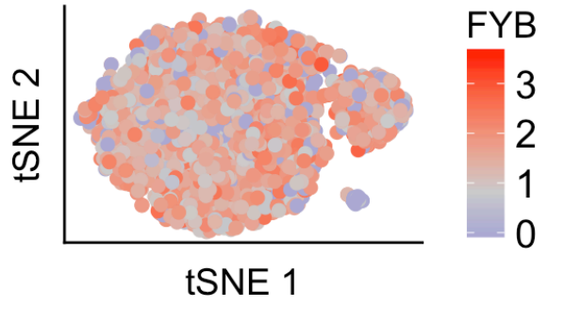


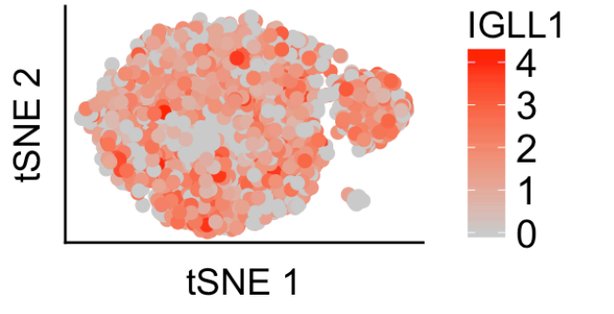


Class 1 - Jurkat cells

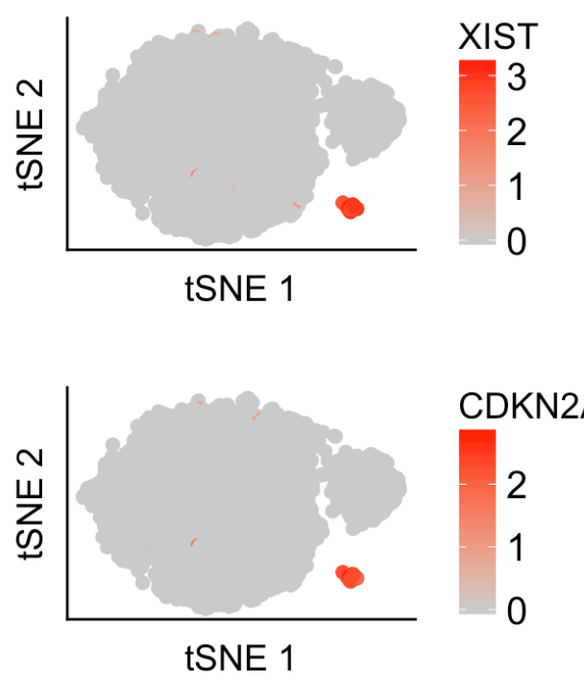


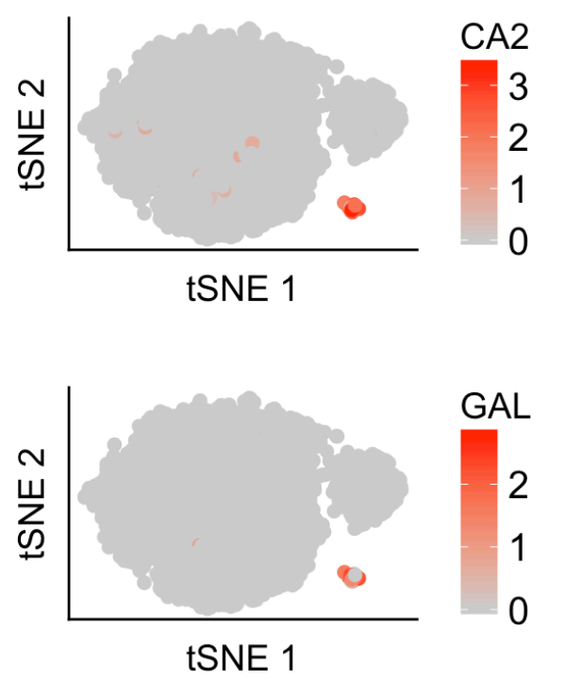


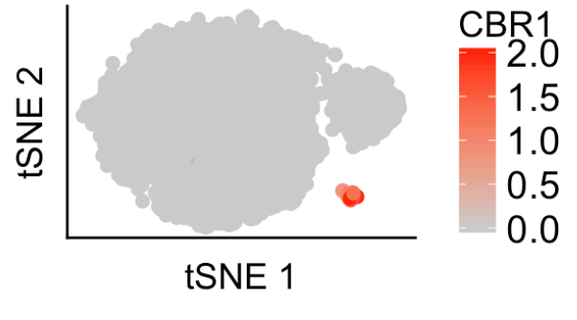


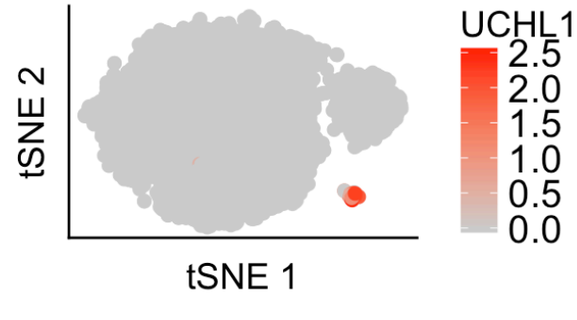


Class 2 - 293T cells



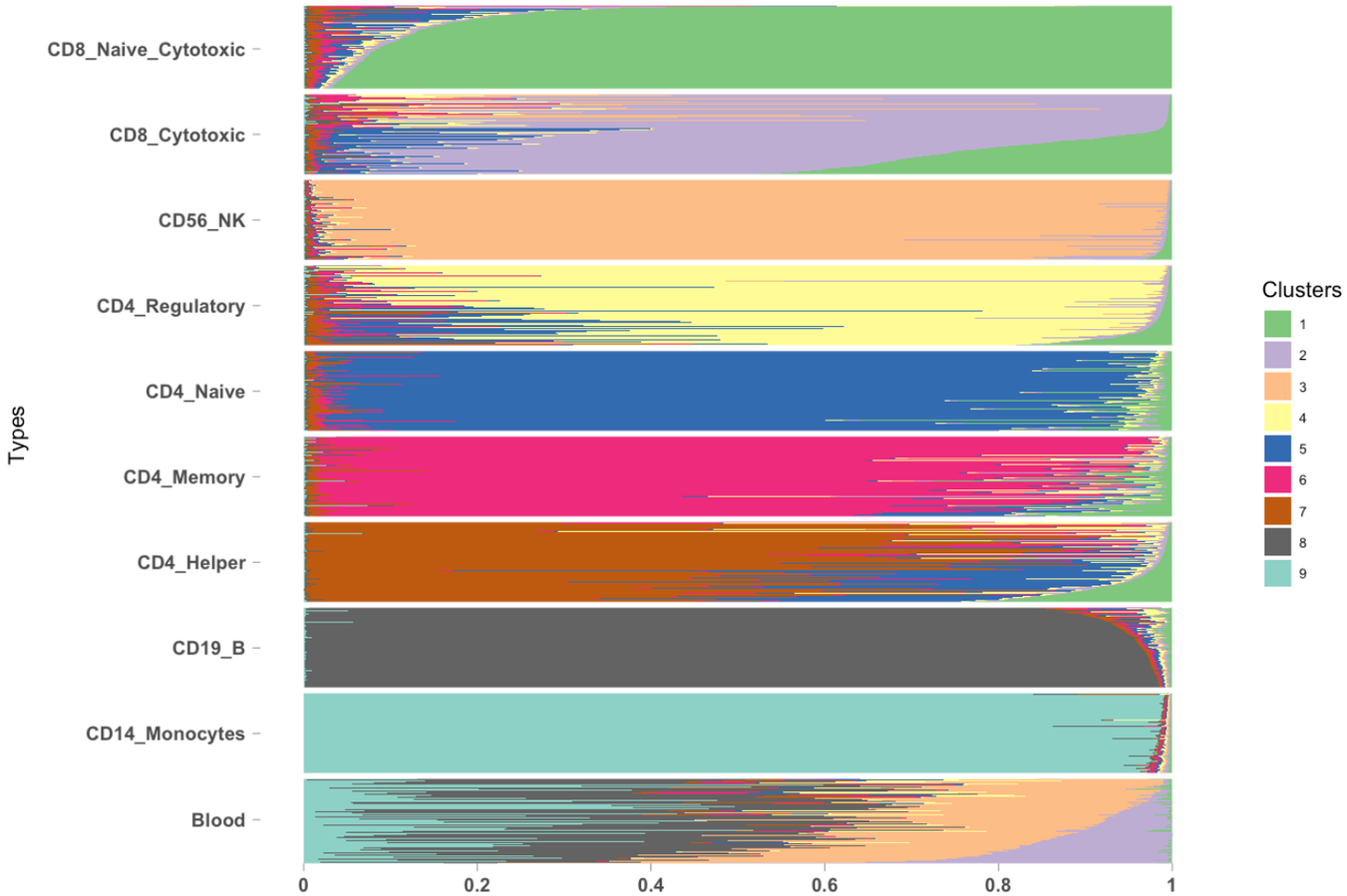




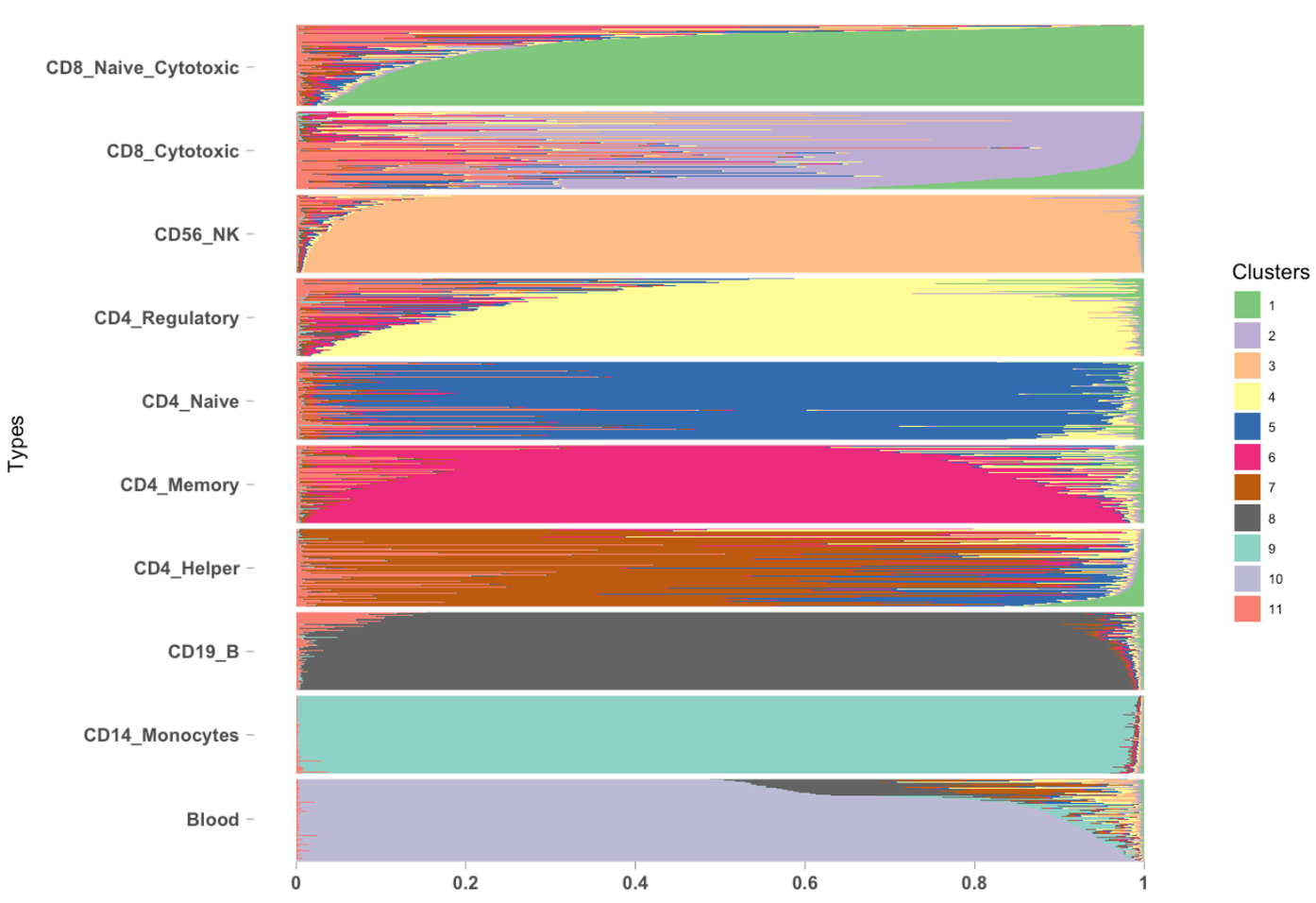


Learning about bulk samples using single cell data

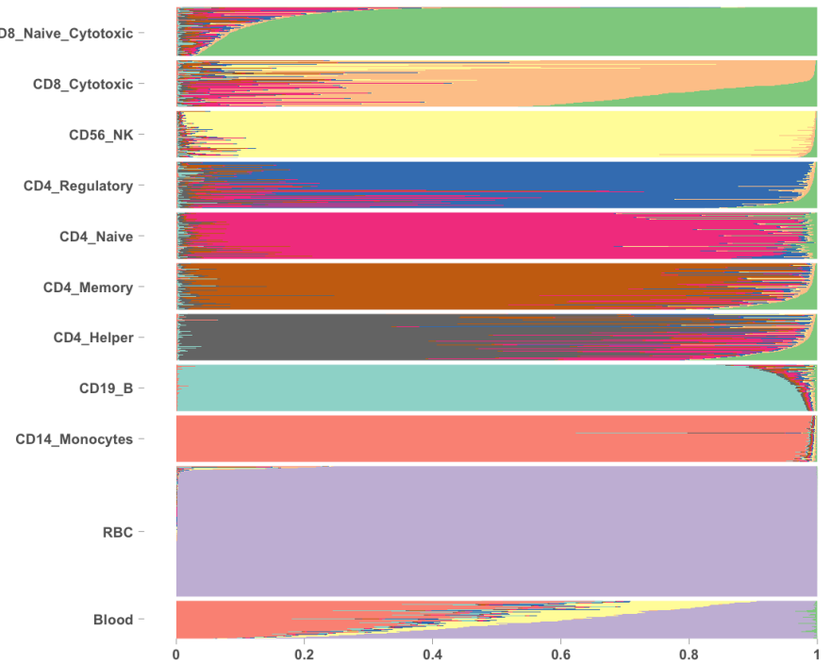
GTEx Whole Blood data +
Zheng et al sorted immune cells

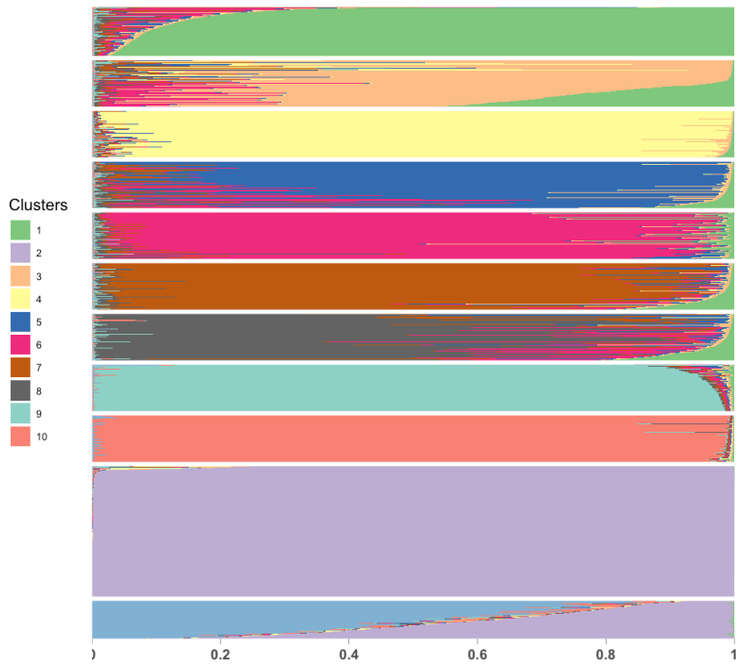


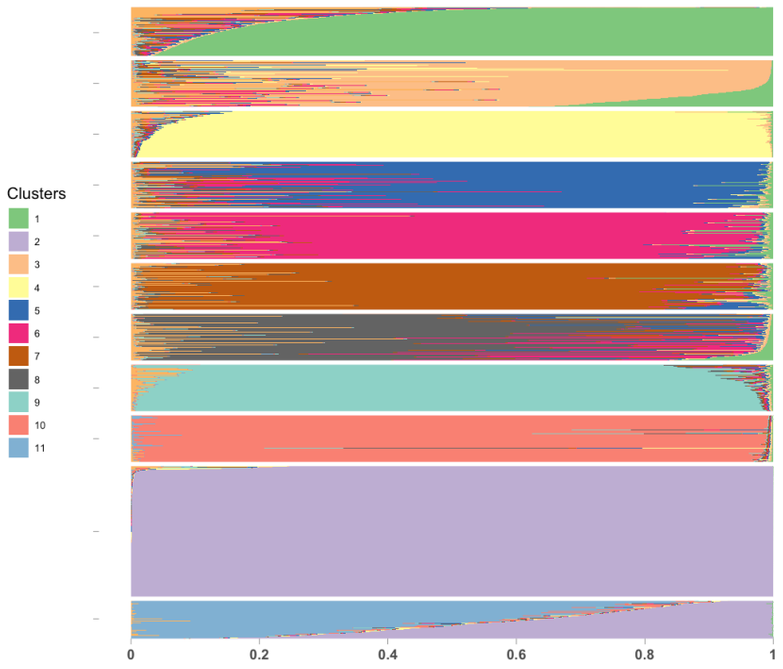
9 cell types
and how they explain
GTEx Whole Blood



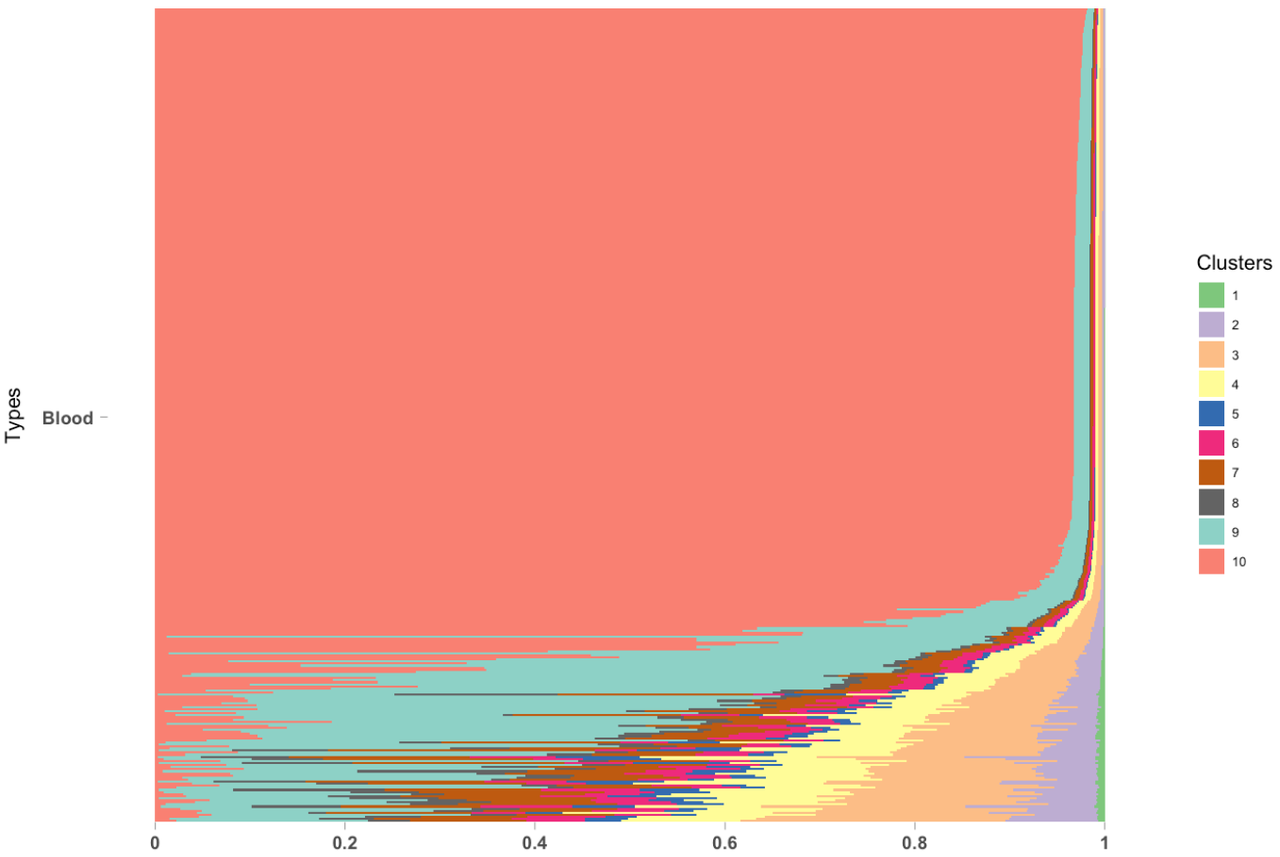
9 cell types + 2 unknown
clusters and how they explain
GTEx Whole Blood



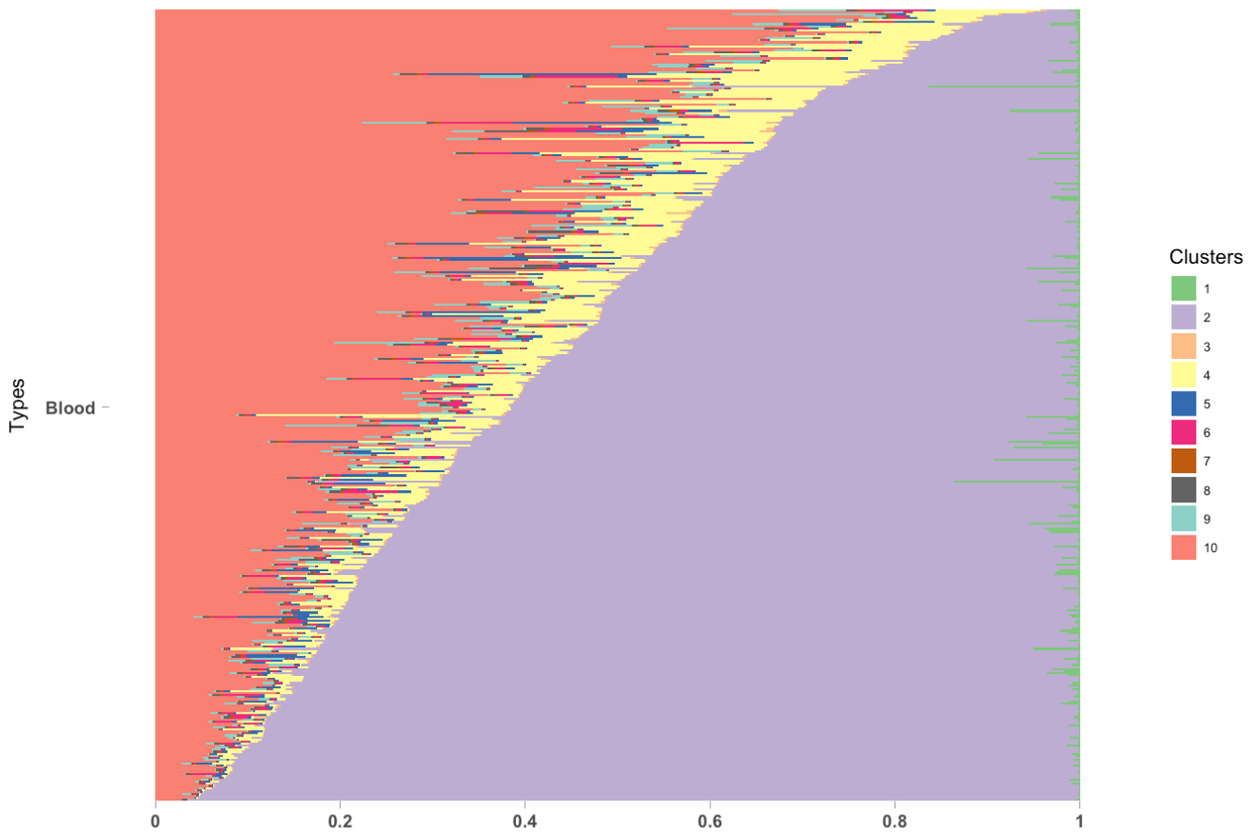




SVM blood



supervised CountClust blood



NOTES

1) unsupervised CountClust apparently seems to be successful in detecting clusters resembling cell types but is not always successful

2) We find a supervised CountClust approach and sometime partially supervised CountClust can often perform better than the unsupervised version.

3) A supervised/partially supervised CountClust allows a better visual representation by t-SNE and can also be used to improve the t-SNE visualization.

4) Supervised CountClust performs better than the SVM on the raw data but slightly worse than the SVM on the normalized data using Seurat normalization. This is likely because supervised CountClust only incorporates read depth normalization.

5) Supervised/Partially supervised CountClust allows to learn the structure of bulk RNA samples based on the training from single cell RNA-seq data, and also allows combining data from multiple sources for training and testing.

Packages : CountClust classtpx

Project Website:

https://kkdey.github.io/singlecell-clustering/

Acknowledgements

Matthew Stephens
Chiaowen Joyce Hsiao
Matt Taddy

Raphael Gottardo
Valentin Voillet
Aude Chapuis
Kelly Paulson
Paul Ngheim