CLUSTERING IN RNA-Seq DATA USING GRADE of MEMBERSHIP MODELS

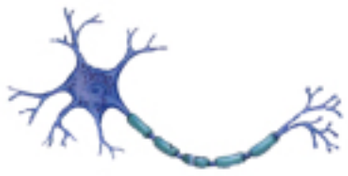
Dey KK, Hsiao CJ, Stephens M (2017) Visualizing the structure of RNA-seq expression data using grade of membership models. PLOS Genetics 13(3): e1006599.https://doi.org/10.1371/journal.pgen.1006599

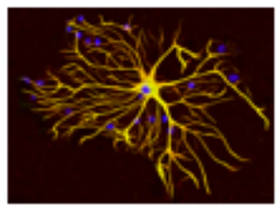
cell type 1

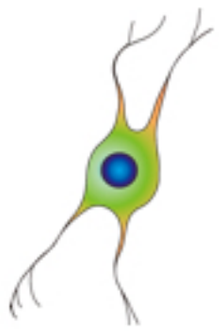
cell type 2

cell type 3

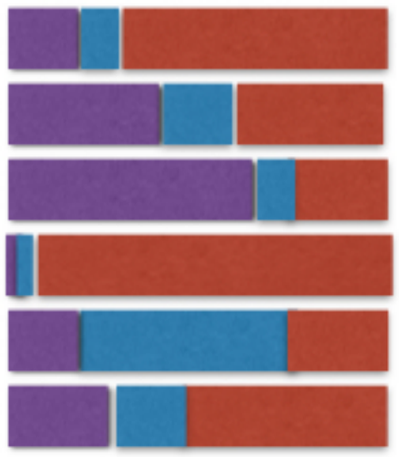
example:
brain



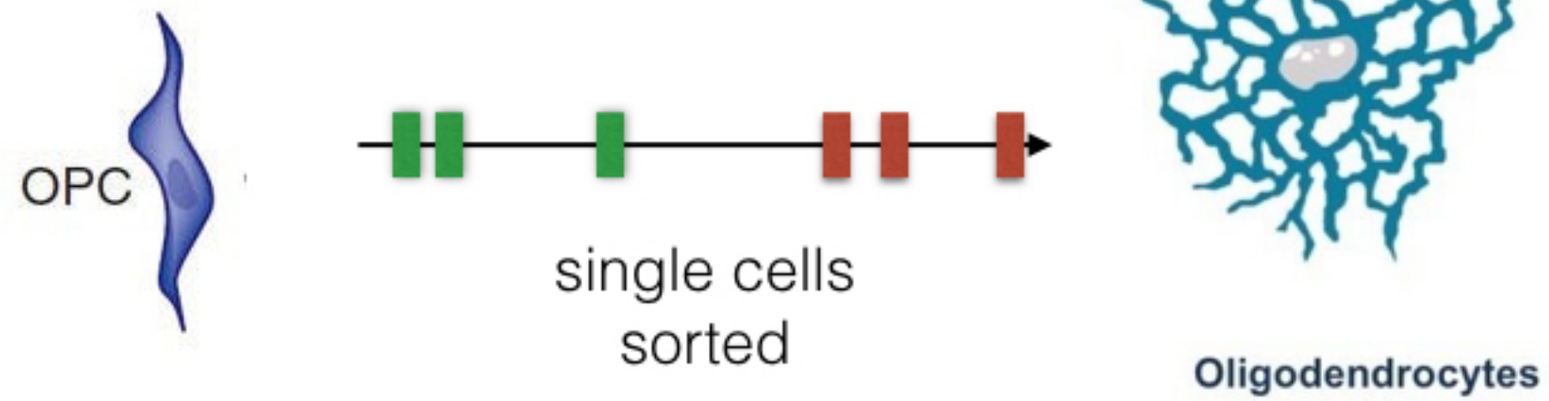




bulk RNA
tissue samples

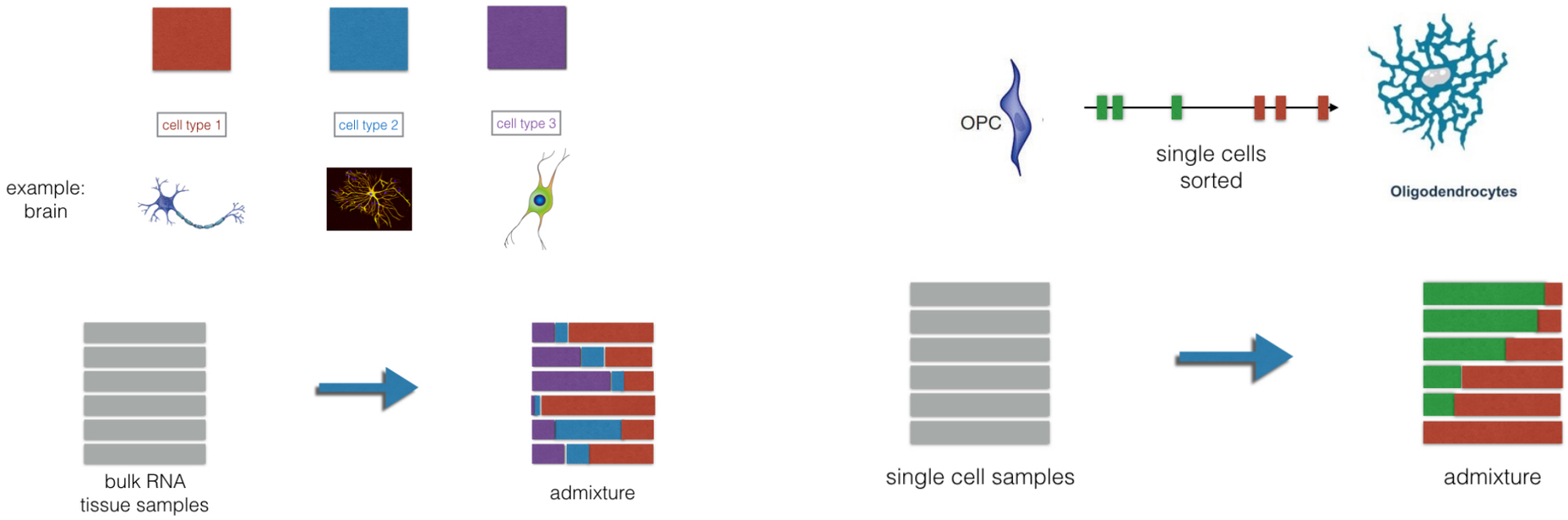


admixture


OPC
single cells
sorted
Oligodendrocytes


single cell samples
admixture

Clustering of RNA-seq data

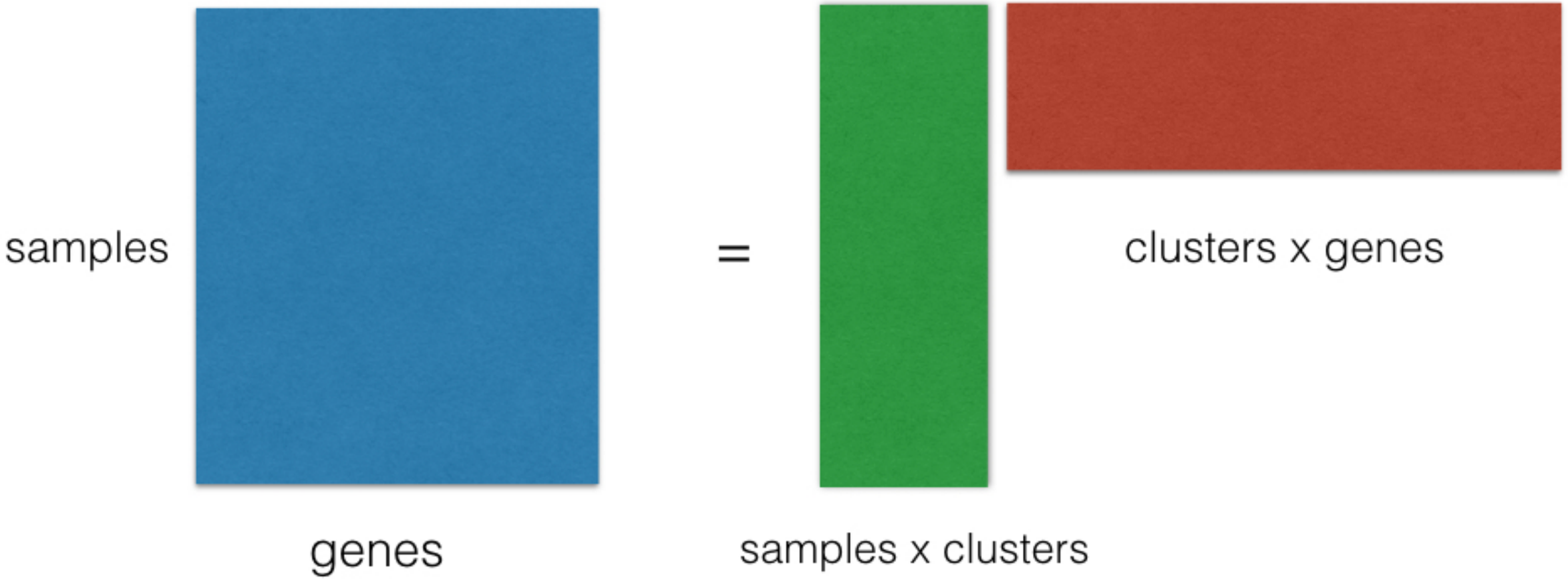


We suggest a model based clustering algorithm in lines of the "admixture" model in population genetics where each sample may have memberships in multiple clusters, which ideally we want to mimic the cell types or cell stages.

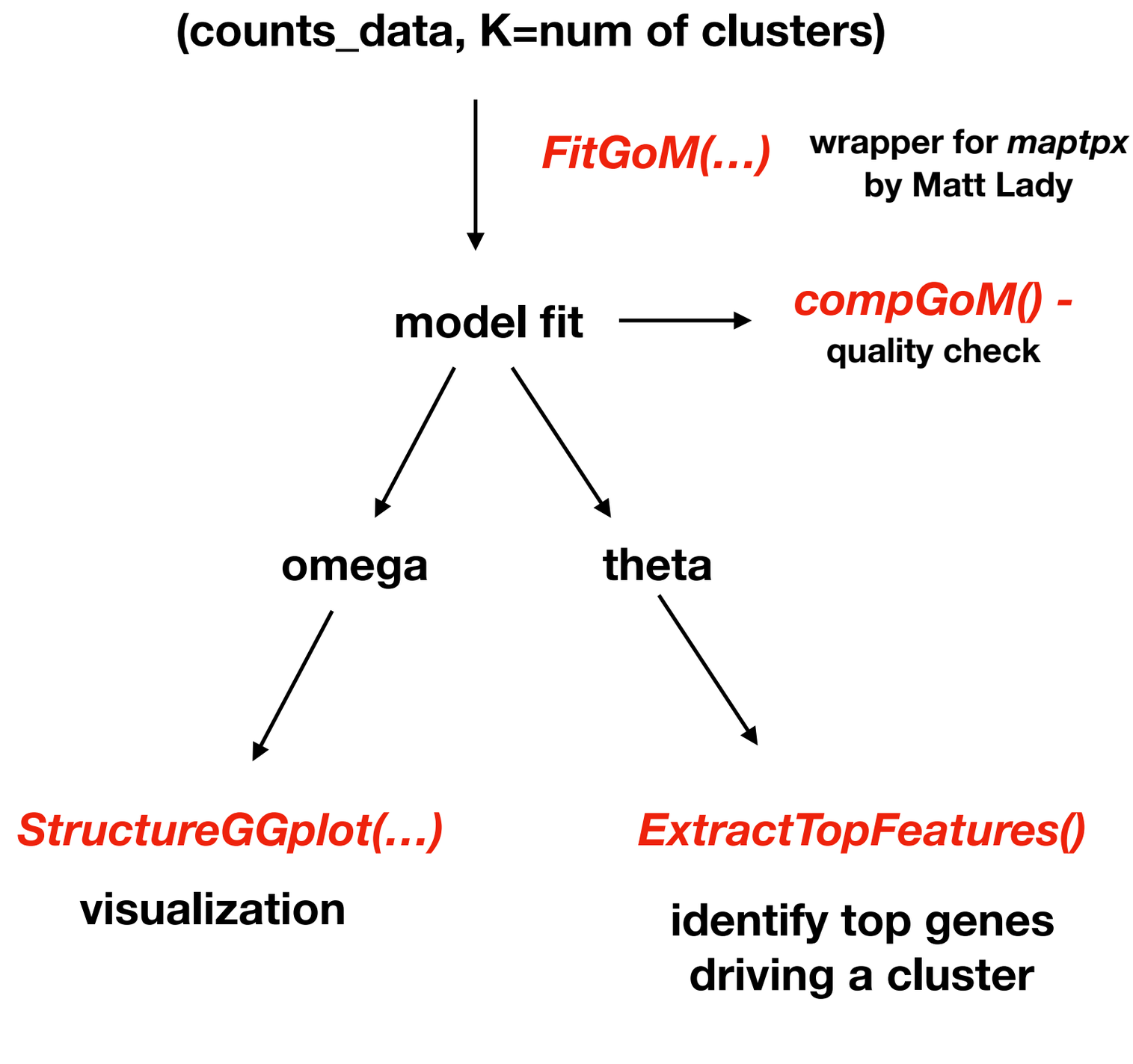
Grade of Membership (GoM) Model

$$(c_{n1}, c_{n2}, \cdots, c_{nG}) \sim Mult(c_{n+}, p_{n1}, p_{n2}, \cdots, p_{nG})$$

$$p_{ng} = \sum_{k=1}^{K} \omega_{nk}\theta_{kg} \qquad \sum_{g=1}^{G} \theta_{kg} = 1 \ \ \forall k \qquad \sum_{k=1}^{K} \omega_{nk} = 1 \ \ \forall n$$



Model workflow (R package CountClust)



How *FitGoM* works?

We assume Dirichlet priors on the parameters ω and θ

both cases with equal mean for all the components

The initial choices of omega and theta are randomly chosen from the prior

An EM type algorithm is run to update the omega and theta until convergence in log posterior probability.

log Bayes factor corresponding to the model and log BIC are reported to check for model fit and compare between model fits (*compGoM*)

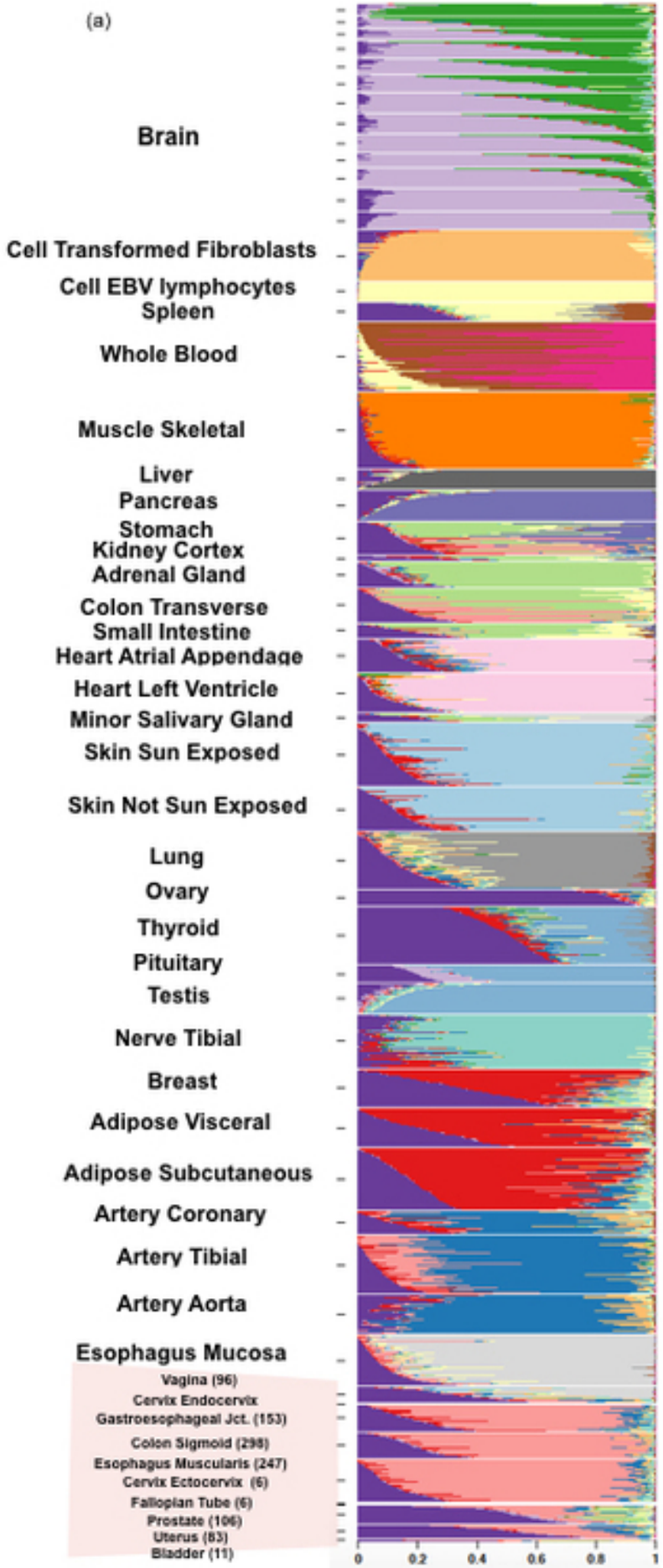
How *ExtractTopFeatures* works?

$$D^g(k) := \min_{l:l\neq k} \left[\theta_{kg} \log \frac{\theta_{kg}}{\theta_{lg}} + \theta_{lg} - \theta_{kg} \right]$$

We compute the above for each gene *g* and cluster *k*

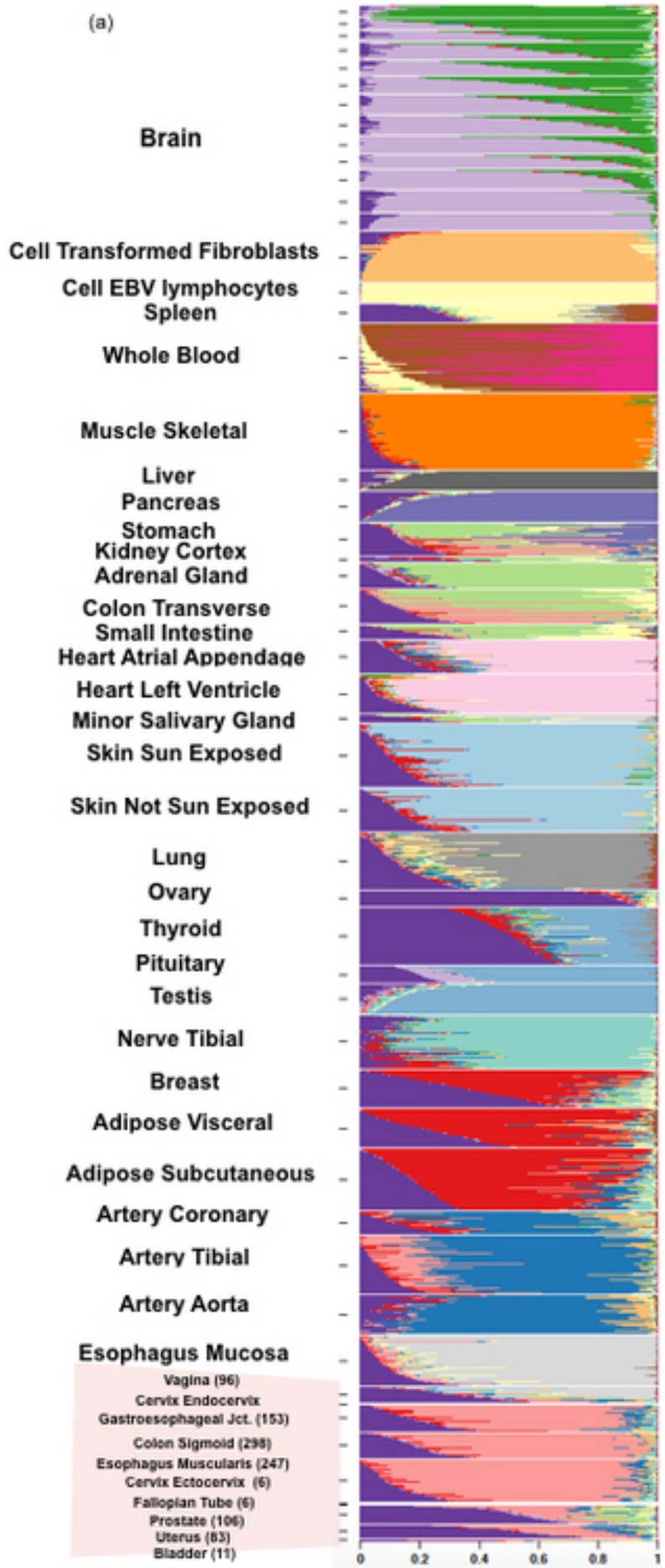
For each k, we identify the genes with highest value of the above measure, and which also have the highest value of θ of all the clusters.

ExtractTopFeatures



Brown cluster : imp genes
CSF3R, MMP25, IL1R2, SELL
Enriched GO: immune response, defense response, cell periphery, regulation of immune response

Pink cluster: imp genes
HBB, HBA1, HBA2, HBD
Enriched GO: hemoglobin complex, heme binding, gas transport

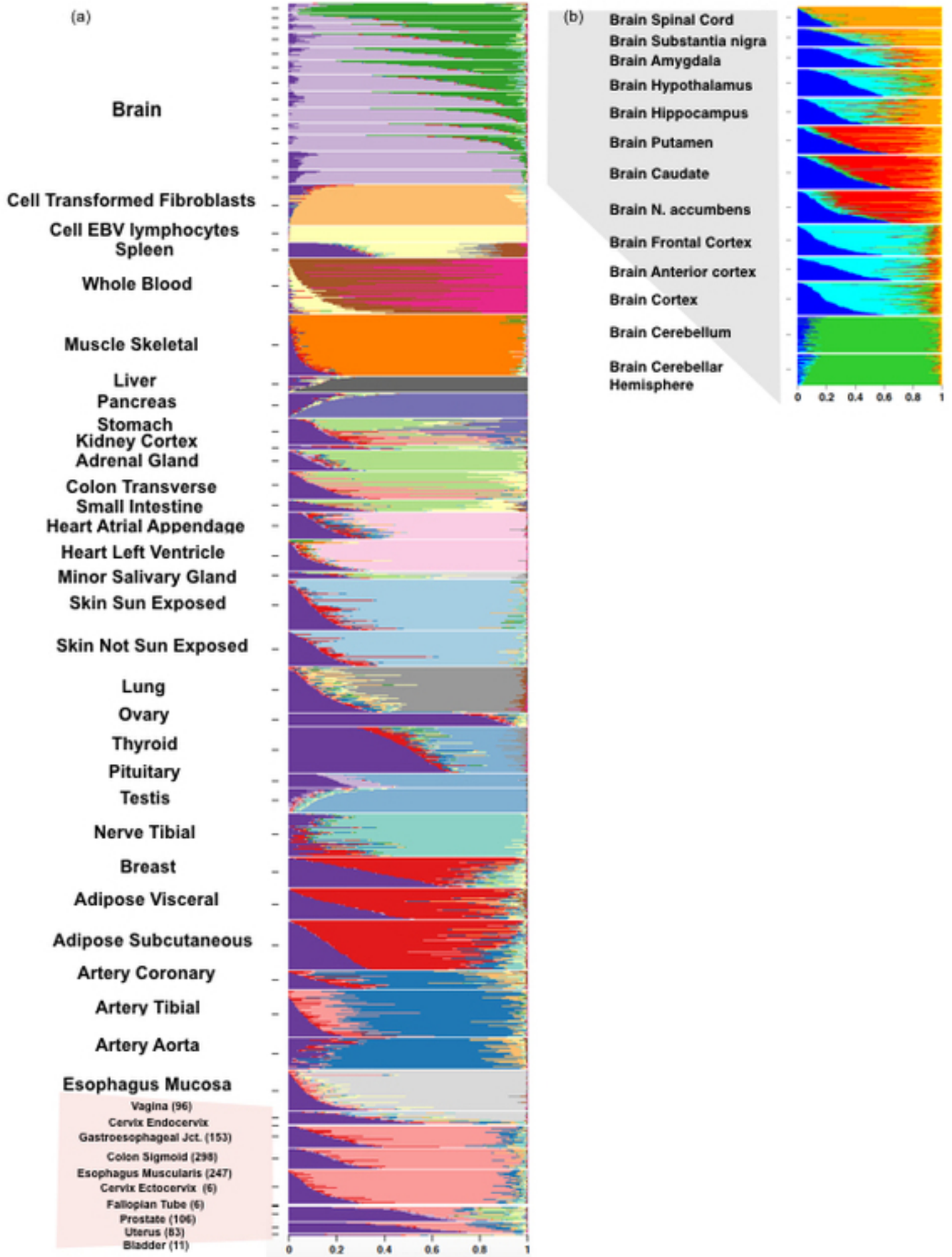


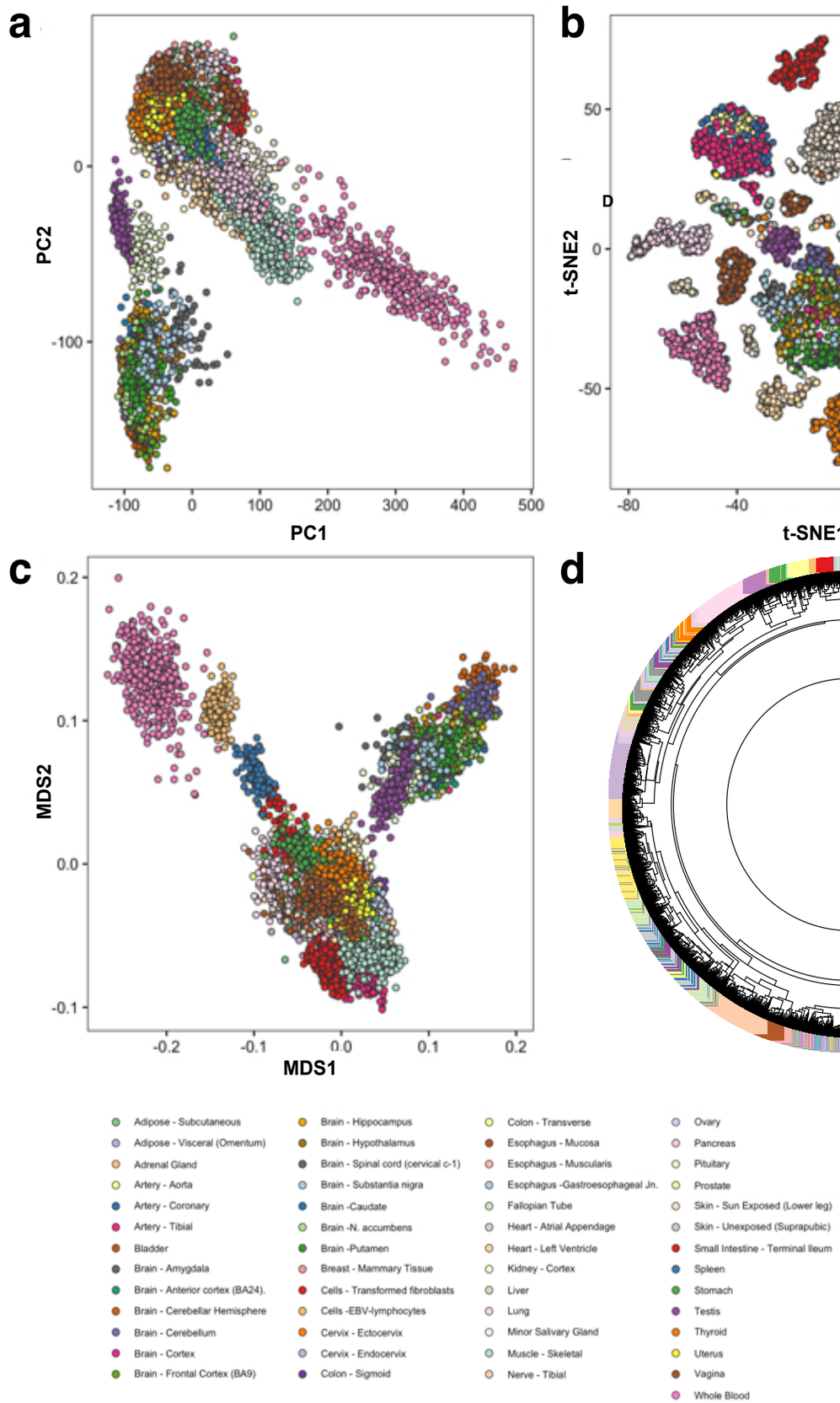
ExtractTopFeatures

Light blue cluster : imp genes
PRSS1, CPA1, PNLIP, CELA3A
Enriched GO: digestion, serine type endopeptidase activity, proteolysis, hydrolase activity

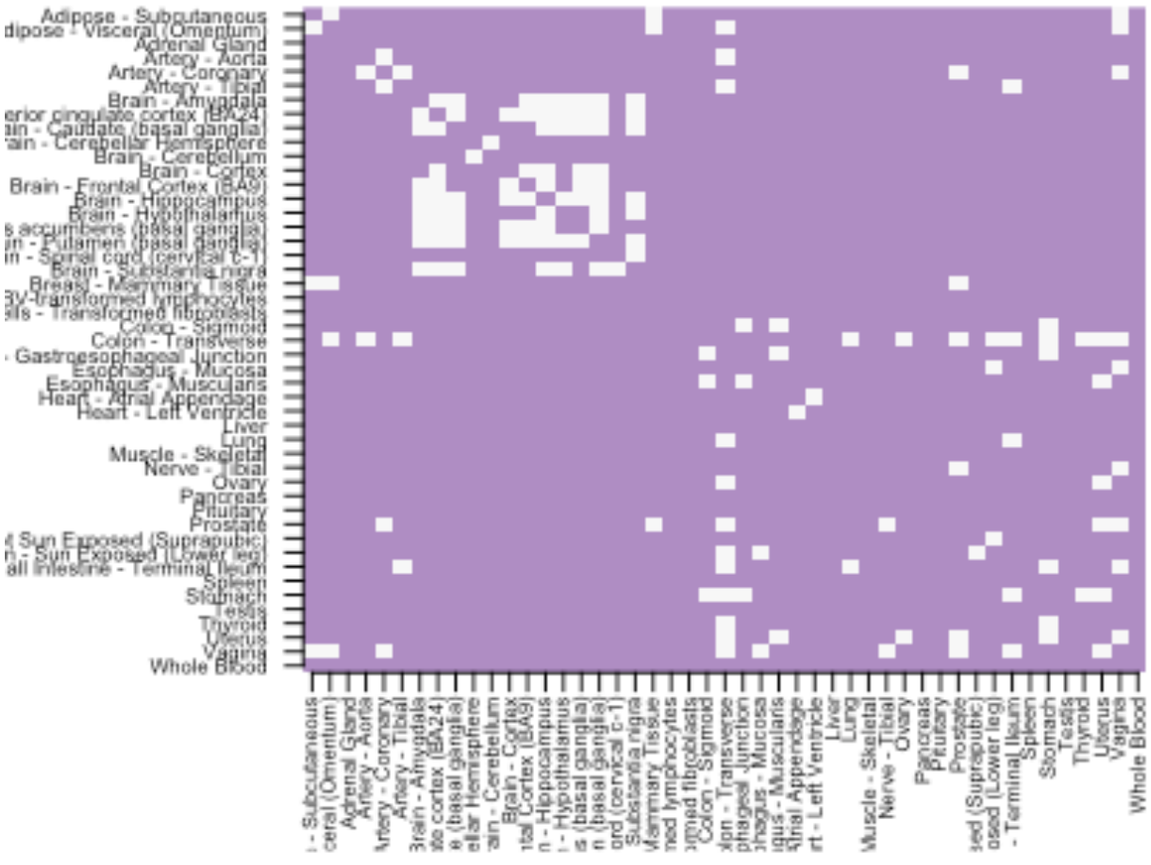
Light blue cluster : imp genes
KRT10, KRT1, KRT2, LOR
Enriched GO: epidermis development, molting cycle, hair cycle

Red cluster: imp genes
FABP4, PLIN1, FASN, GPX3
Enriched GO: cellular lipid metabolism, acylglycerol metabolism, angiogenesis regulation

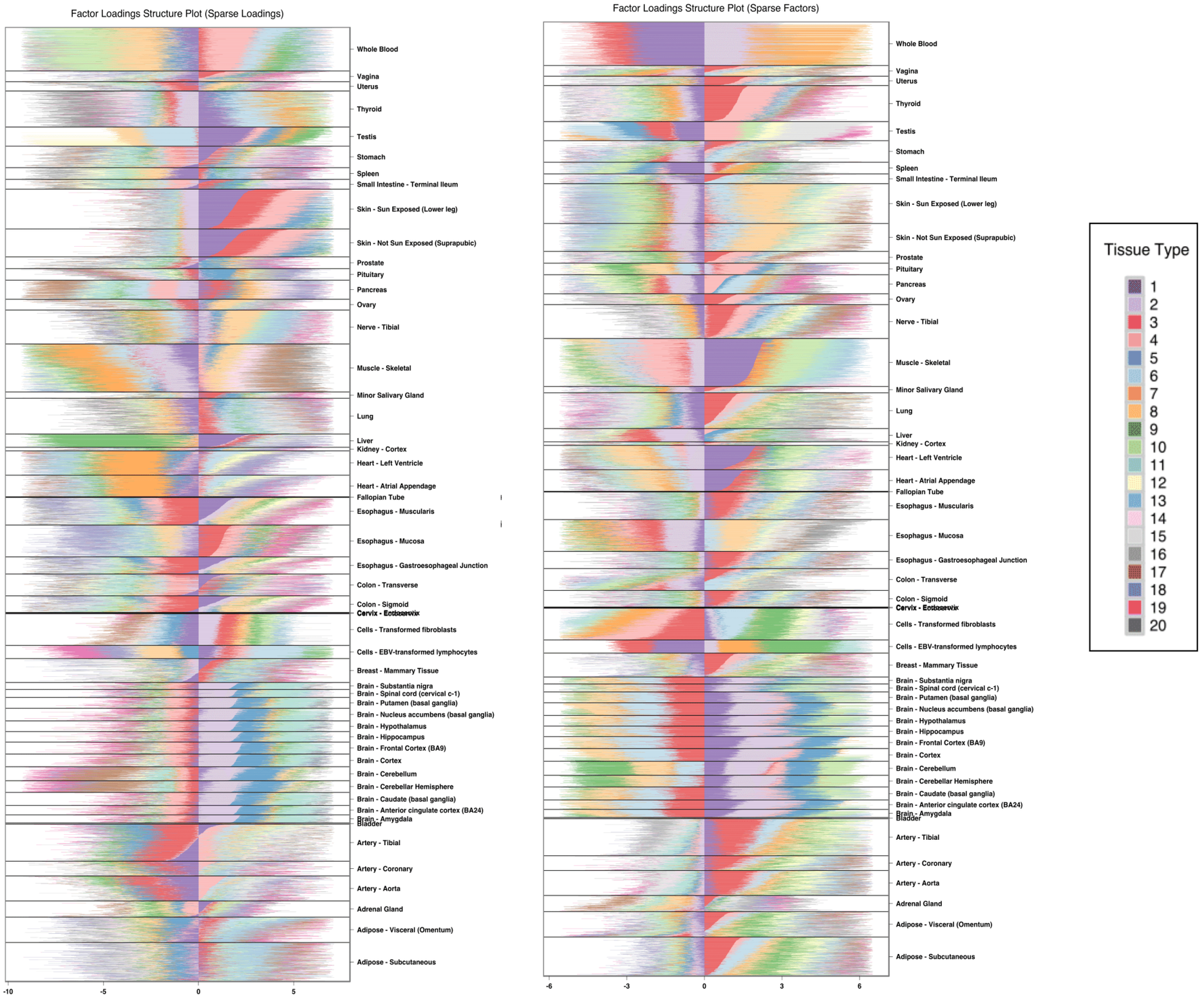

(a)
Brain
Cell Transformed Fibroblasts
Cell EBV lymphocytes
Spleen
Whole Blood
Muscle Skeletal
Liver
Pancreas
Stomach
Kidney Cortex
Adrenal Gland
Colon Transverse
Small Intestine
Heart Atrial Appendage
Heart Left Ventricle
Minor Salivary Gland
Skin Sun Exposed
Skin Not Sun Exposed
Lung
Ovary
Thyroid
Pituitary
Testis
Nerve Tibial
Breast
Adipose Visceral
Adipose Subcutaneous
Artery Coronary
Artery Tibial
Artery Aorta
Esophagus Mucosa
Vagina (96)
Cervix Endocervix
Gastroesophageal Jct. (153)
Colon Sigmoid (298)
Esophagus Muscularis (247)
Cervix Ectocervix (6)
Fallopian Tube (6)
Prostate (106)
Uterus (83)
Bladder (11)
0
0.2
0.4
0.6
0.8
1
(b)
Brain Spinal Cord
Brain Substantia nigra
Brain Amygdala
Brain Hypothalamus
Brain Hippocampus
Brain Putamen
Brain Caudate
Brain N. accumbens
Brain Frontal Cortex
Brain Anterior cortex
Brain Cortex
Brain Cerebellum
Brain Cerebellar Hemisphere
0
0.2
0.4
0.6
0.8
1


a
PC2
PC1
b
t-SNE2
t-SNE1
c
MDS2
MDS1
d
Adipose - Subcutaneous
Adipose - Visceral (Omentum)
Adrenal Gland
Artery - Aorta
Artery - Coronary
Artery - Tibial
Bladder
Brain - Amygdala
Brain - Anterior cortex (BA24).
Brain - Cerebellar Hemisphere
Brain - Cerebellum
Brain - Cortex
Brain - Frontal Cortex (BA9)
Brain - Hippocampus
Brain - Hypothalamus
Brain - Spinal cord (cervical c-1)
Brain - Substantia nigra
Brain -Caudate
Brain -N. accumbens
Brain -Putamen
Breast - Mammary Tissue
Cells - Transformed fibroblasts
Cells -EBV-lymphocytes
Cervix - Ectocervix
Cervix - Endocervix
Colon - Sigmoid
Colon - Transverse
Esophagus - Mucosa
Esophagus - Muscularis
Esophagus -Gastroesophageal Jn.
Fallopian Tube
Heart - Atrial Appendage
Heart - Left Ventricle
Kidney - Cortex
Liver
Lung
Minor Salivary Gland
Muscle - Skeletal
Nerve - Tibial
Ovary
Pancreas
Pituitary
Prostate
Skin - Sun Exposed (Lower leg)
Skin - Unexposed (Suprapubic)
Small Intestine - Terminal Ileum
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
Whole Blood

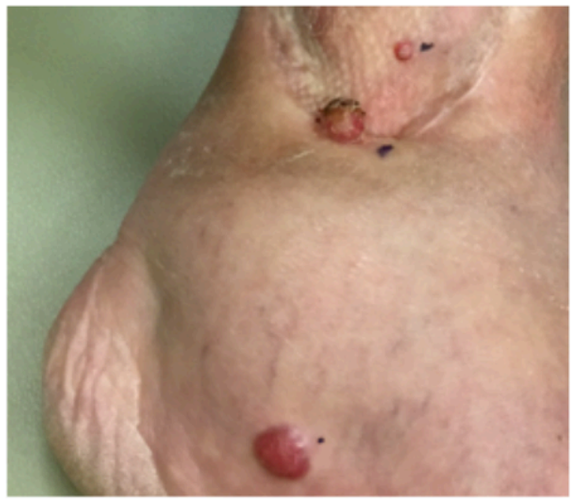
Comparison - GoM vs Hierarchical Method



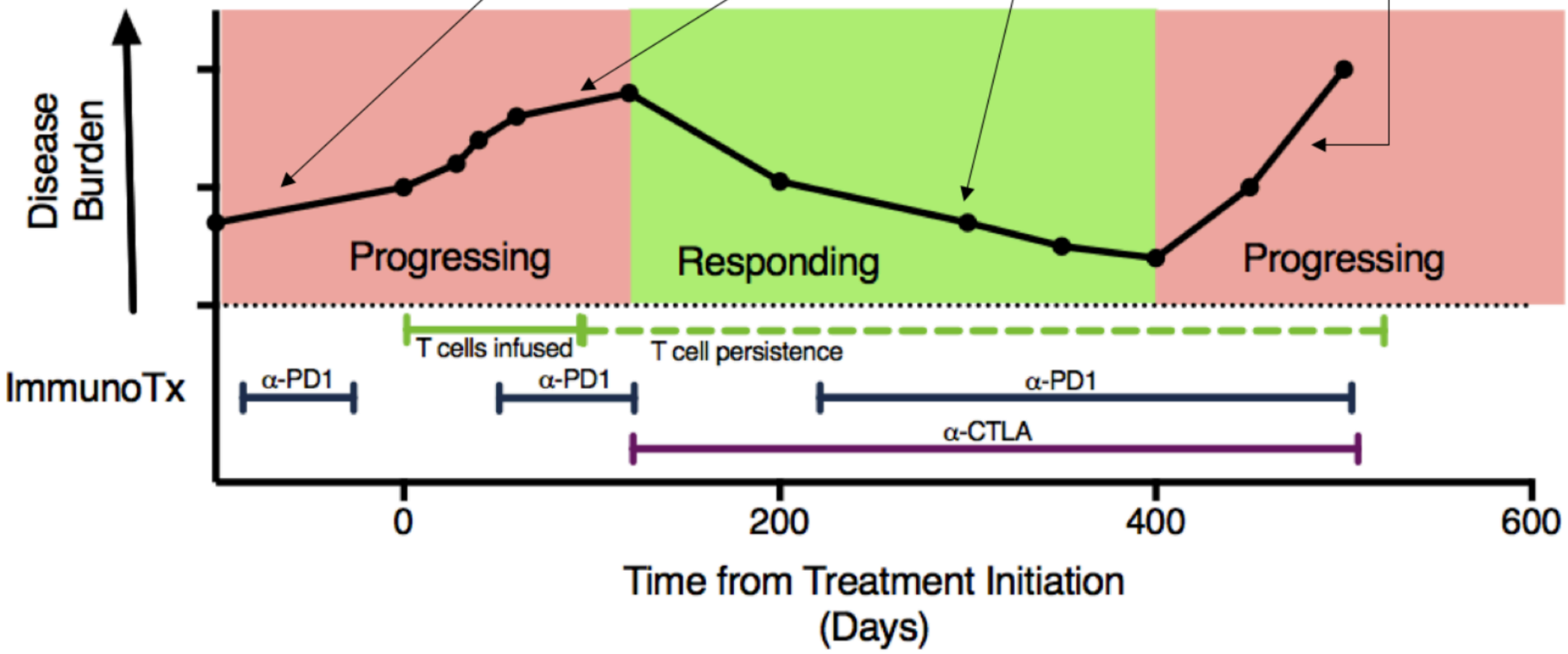
Sparse Factor Analysis : Engelhardt and Stephens (2010)

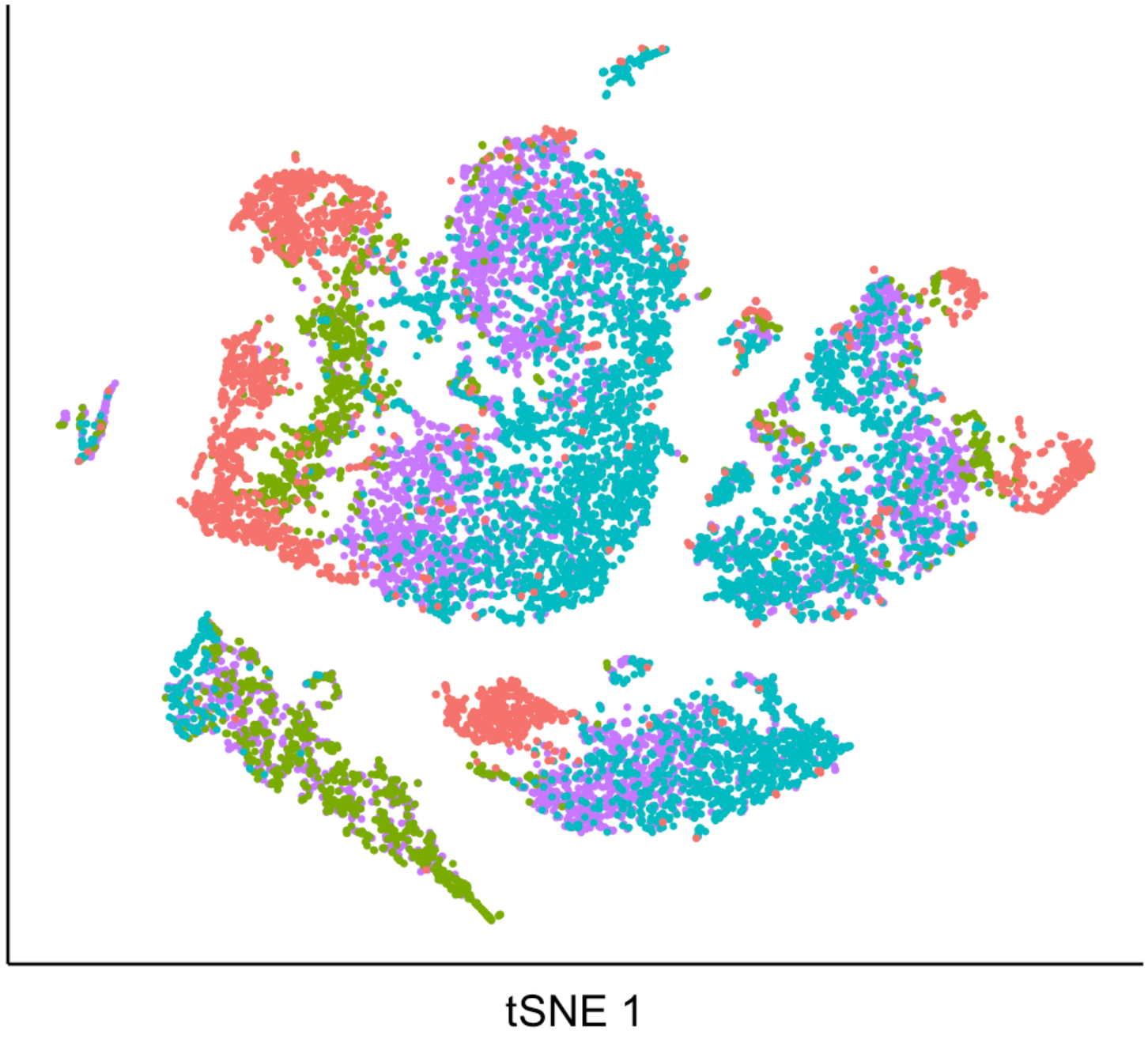


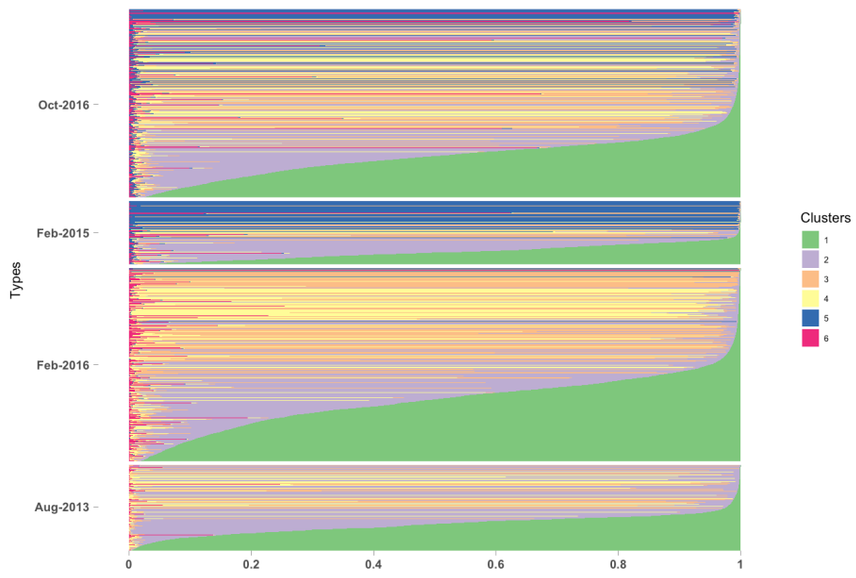
Merkel Cell Carcinoma Study

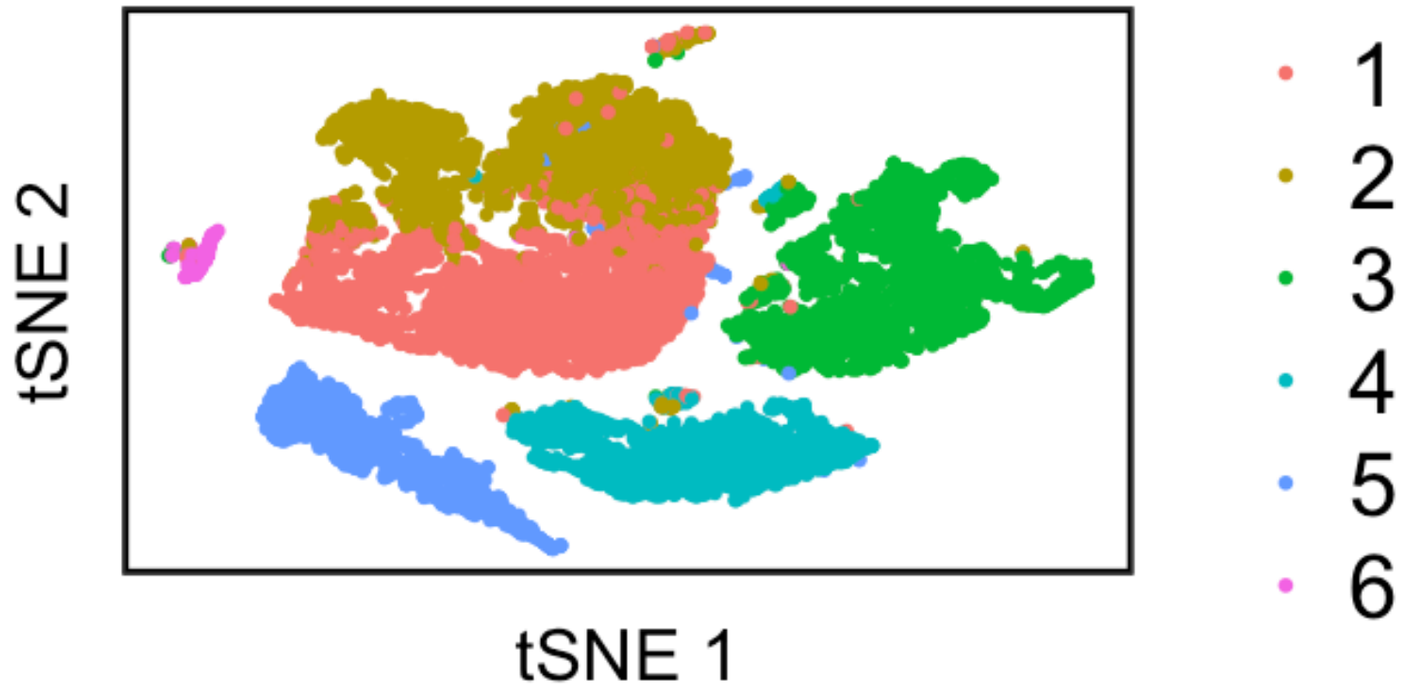


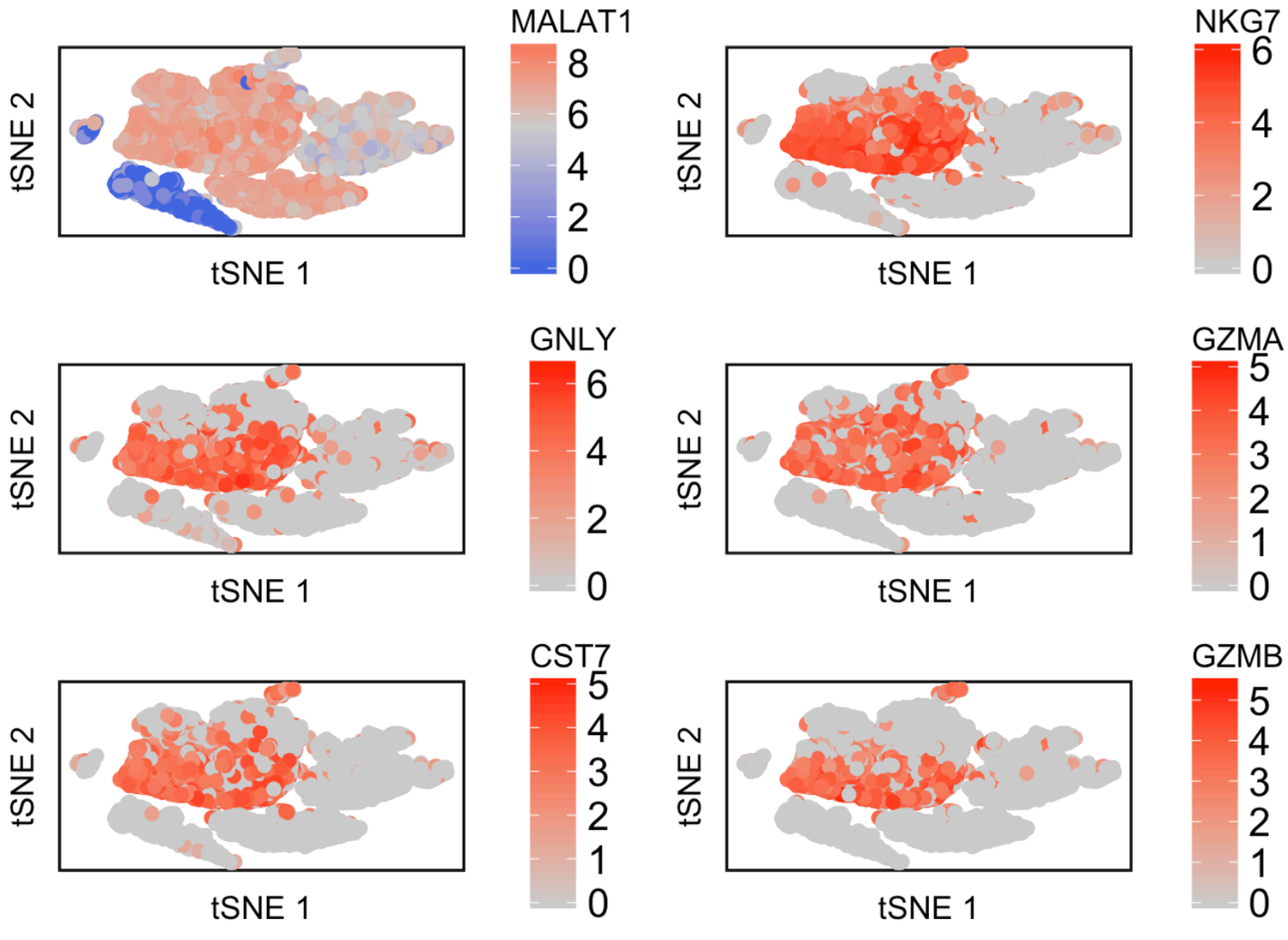
	Aug 2013	Feb 2015	Feb 2016	Oct 2016
Time point:	**Pre-ImmunoTx**	**Failing T cells & α-PD-1**	**Responding to T cells & α-PD-1 & α-CTLA**	**Failing T cells & α-PD-1 & α-CTLA**
Samples for 10X-Genomics	Single-cell tumor digest (viable) TIL (viable) PBMC	PBMC	PBMC	Single-cell tumor digest (viable) TIL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




tSNE 2
tSNE 1
Aug-2013
Feb-2015
Feb-2016
Oct-2016

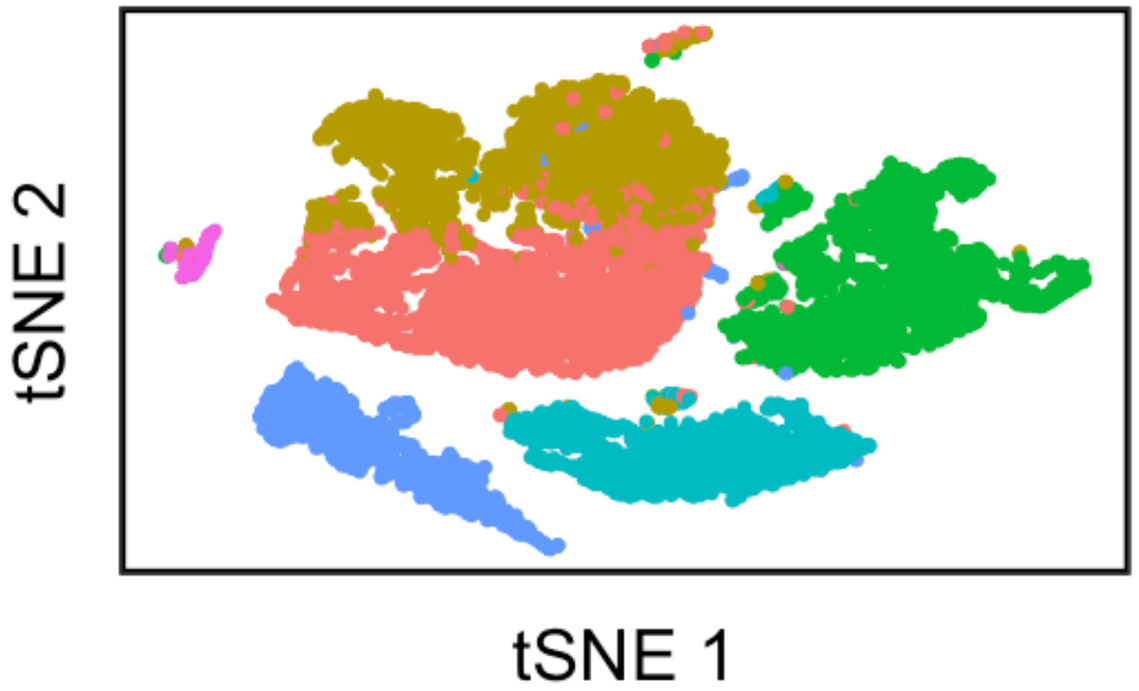

Oct-2016
Feb-2015
Feb-2016
Aug-2013
Types
0
0.2
0.4
0.6
0.8
1
Clusters
1
2
3
4
5
6

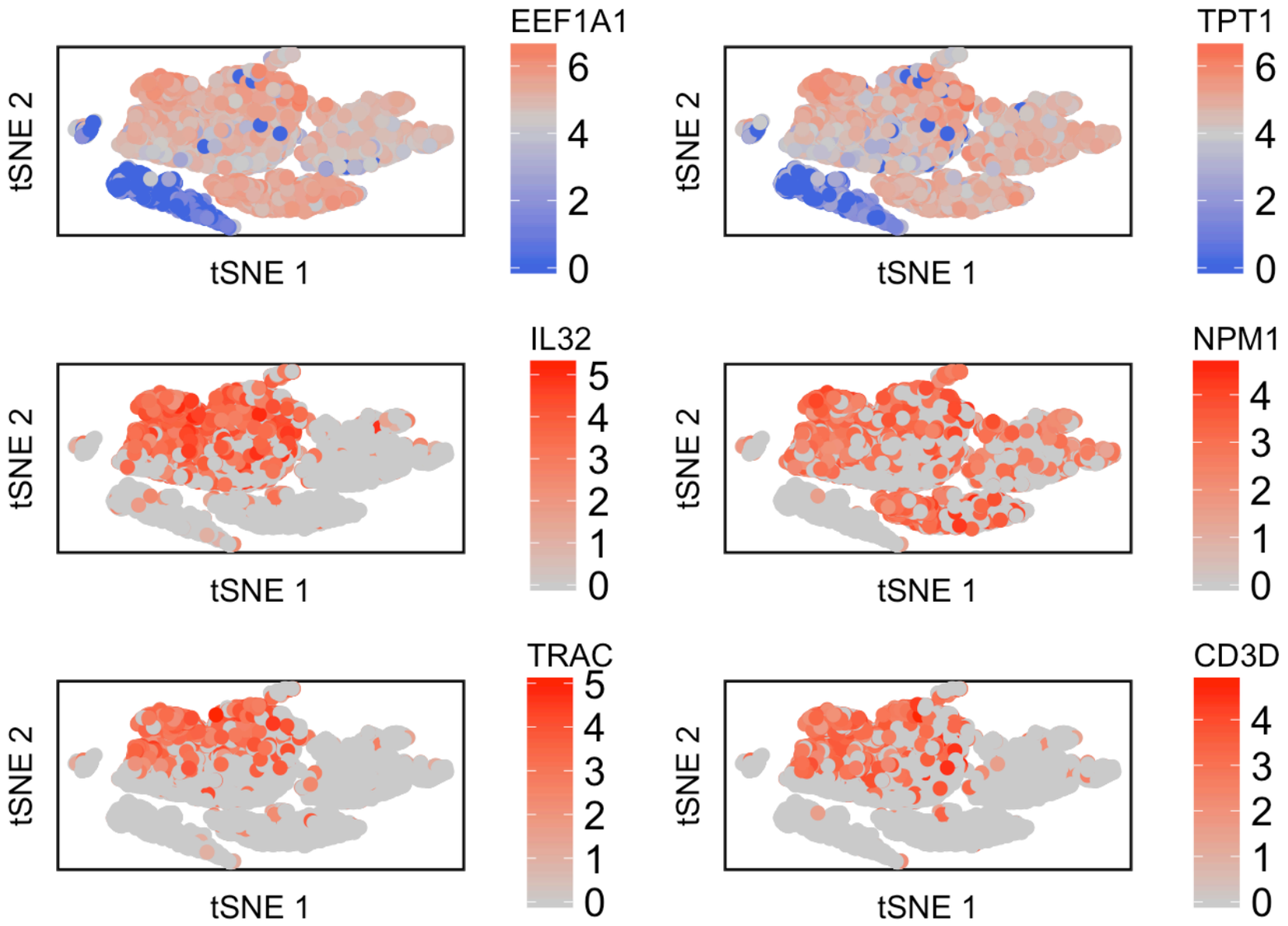




Top enriched pathways

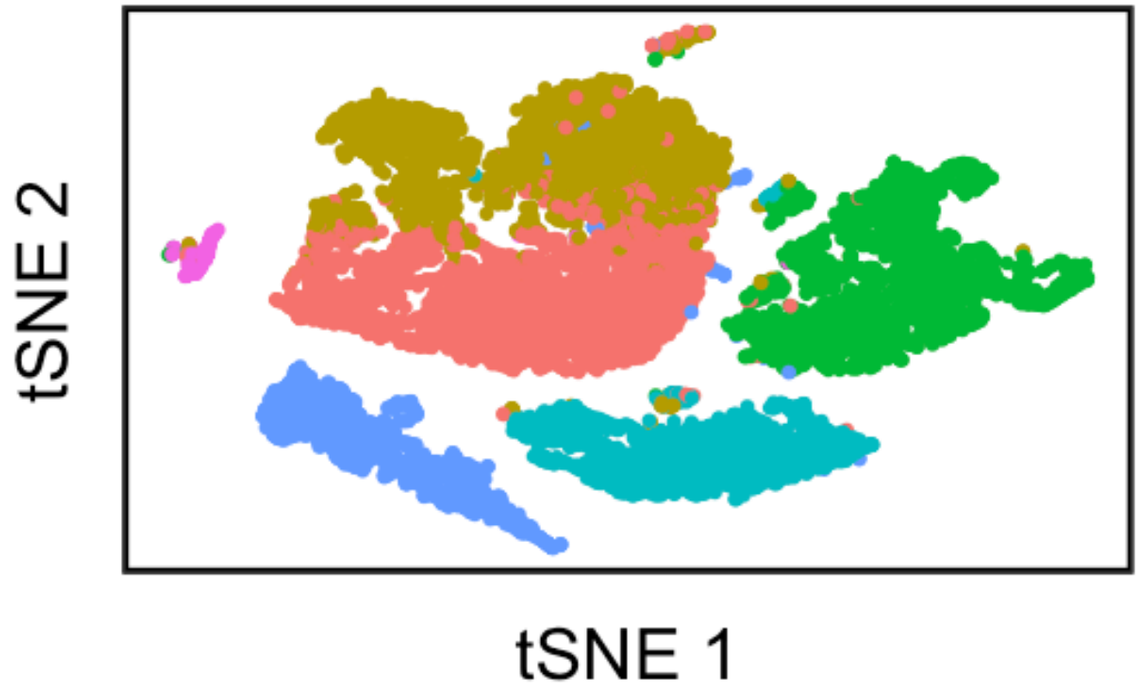
1. Natural killer cell mediated cytotoxicity - Homo sapiens (human)
2. Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell
3. Immune System
4. IL12-mediated signaling events
5. Downstream signaling in naïve CD8+ T cells

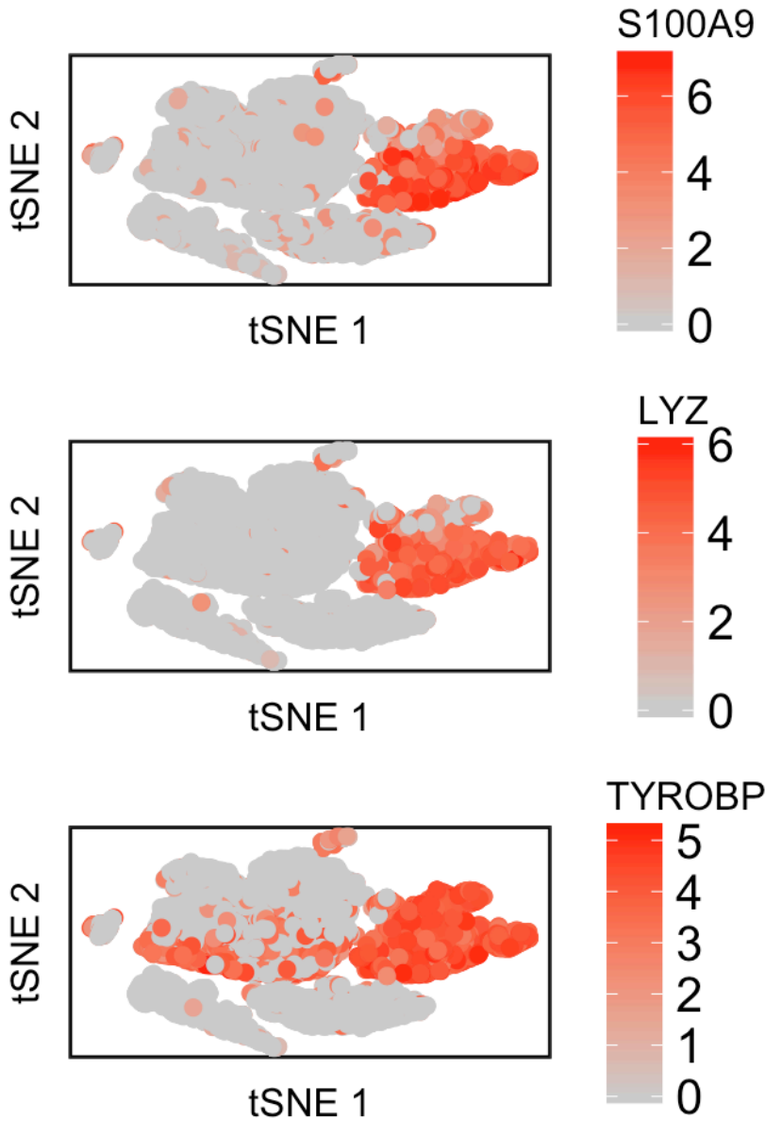


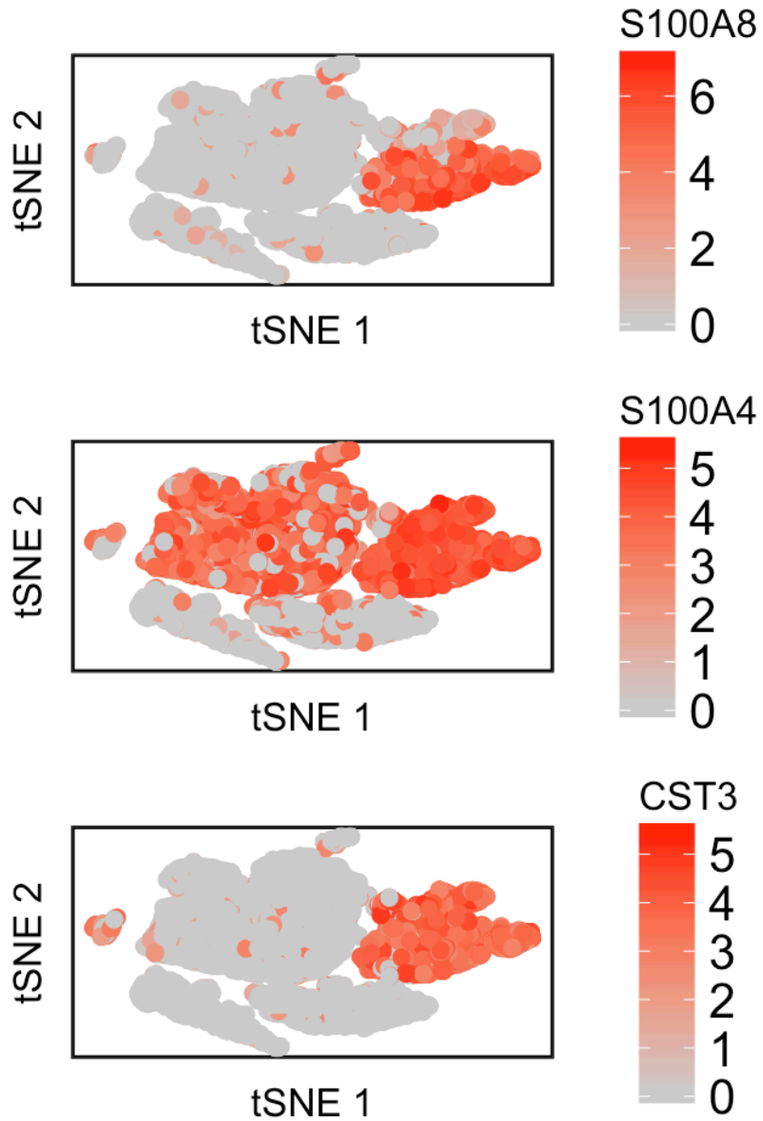


Top enriched pathways

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

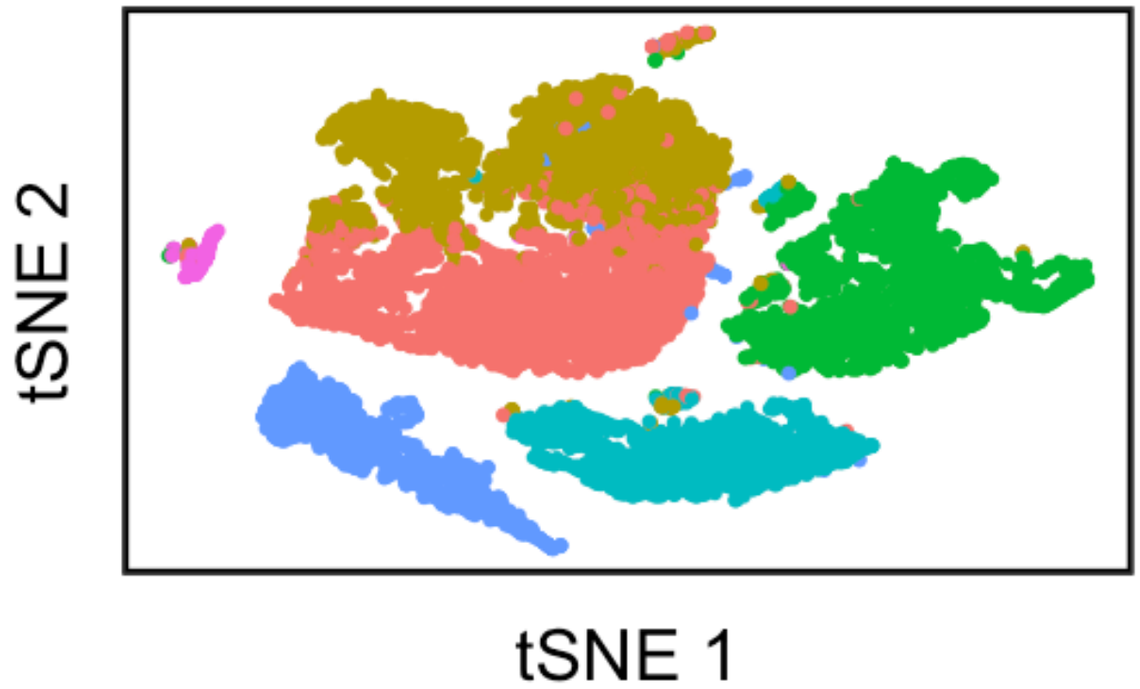


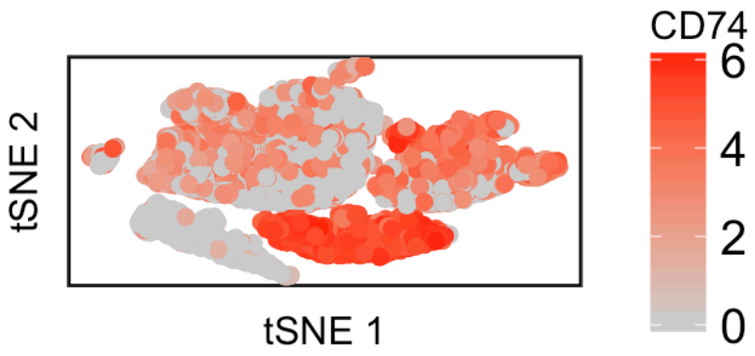


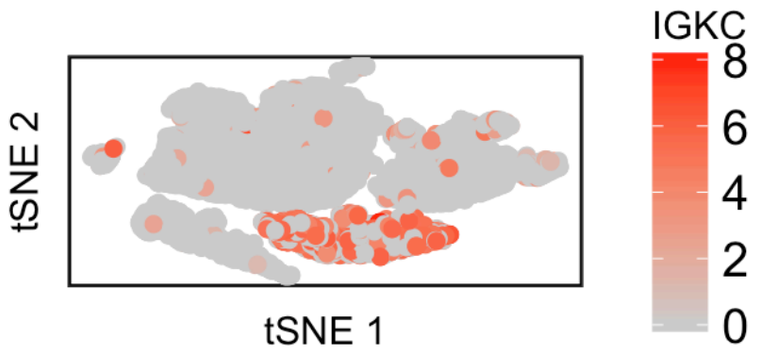


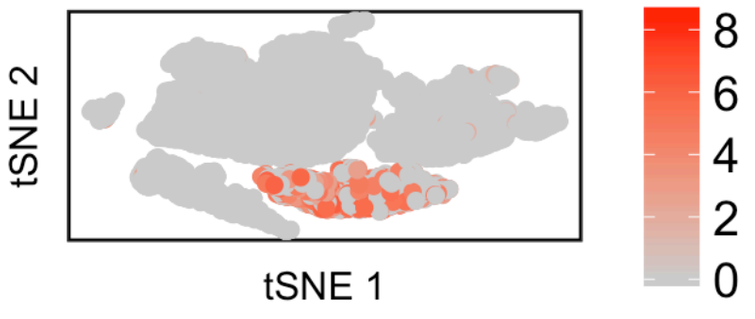
Top enriched pathways

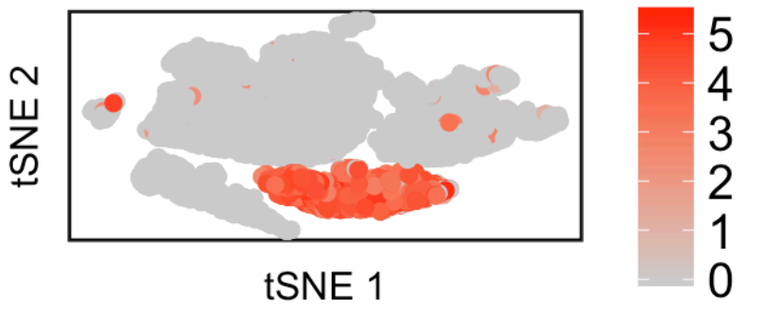
1. Neutrophil degranulation
2. Innate Immune System
3. Immune System
4. Toll-Like Receptors Cascades
5. Endogenous TLR signaling
6. Regulation of TLR by endogenous ligand

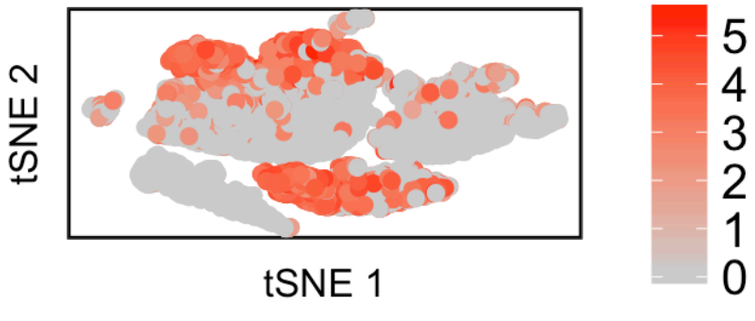


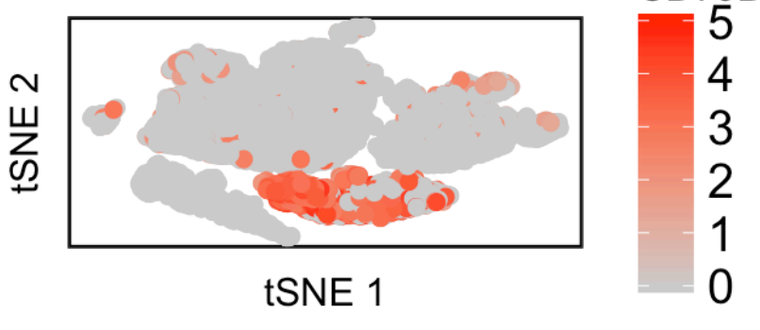






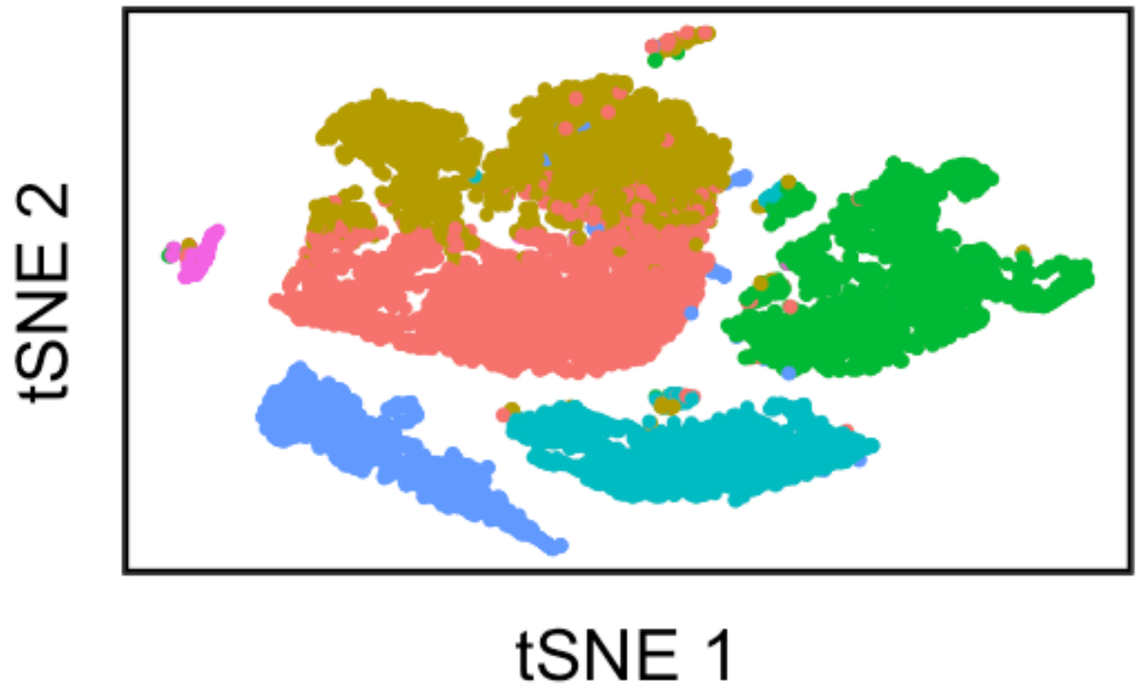


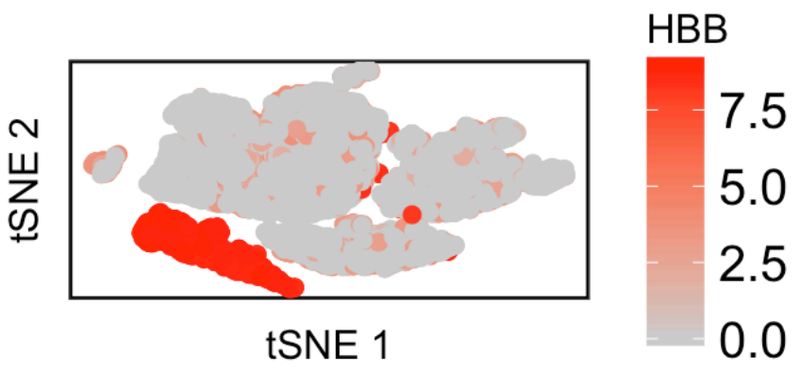


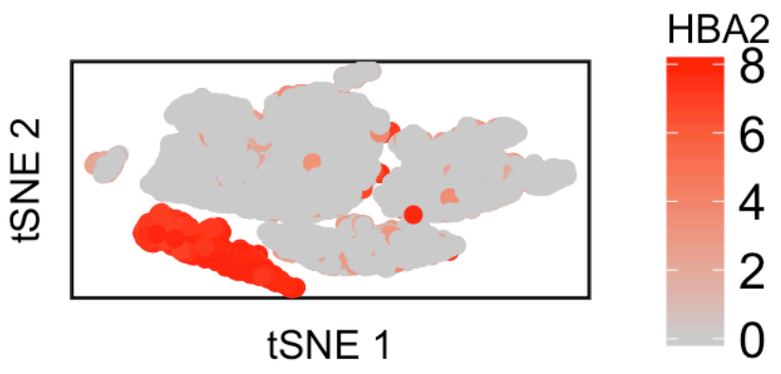


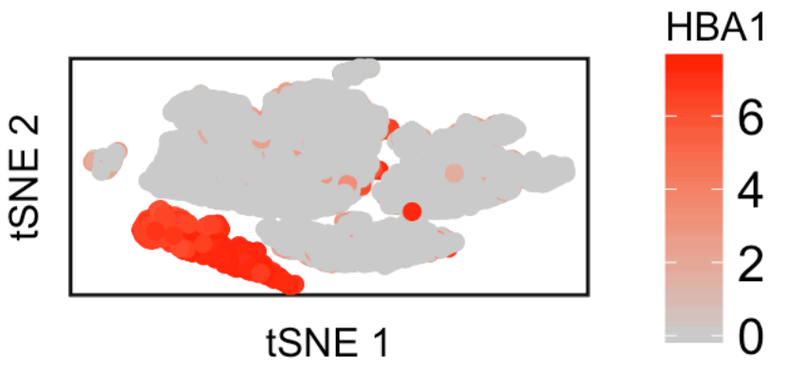
Top enriched pathways

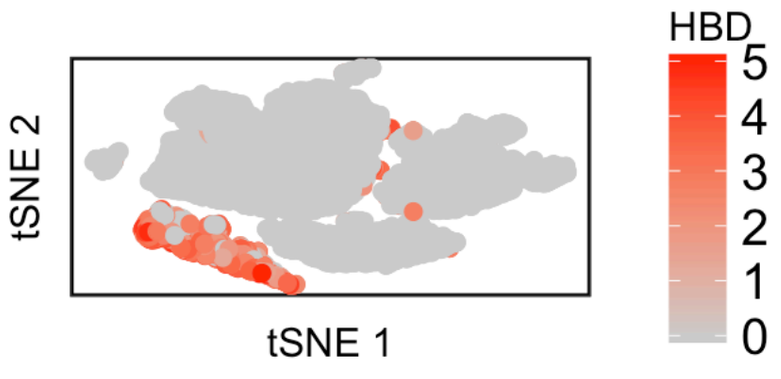
1. CD22 mediated BCR regulation
2. Antigen activates B Cell Receptor (BCR) leading to generation of second messengers
3. Cell surface interactions at the vascular wall
4. BCR
5. Signaling by the B Cell Receptor (BCR)

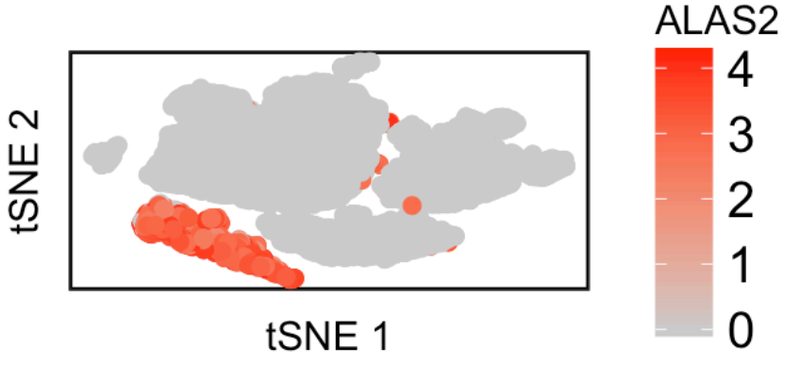


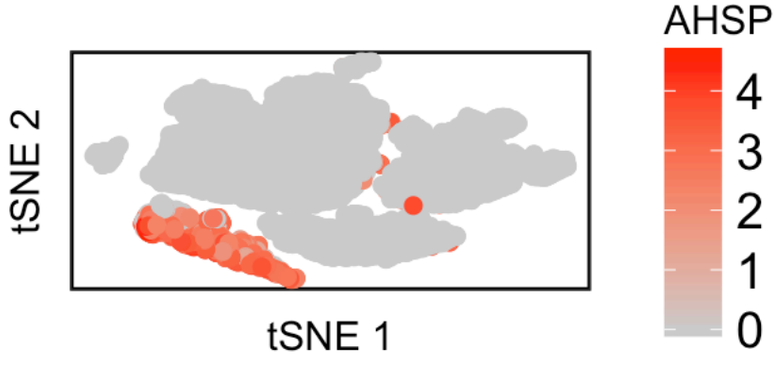






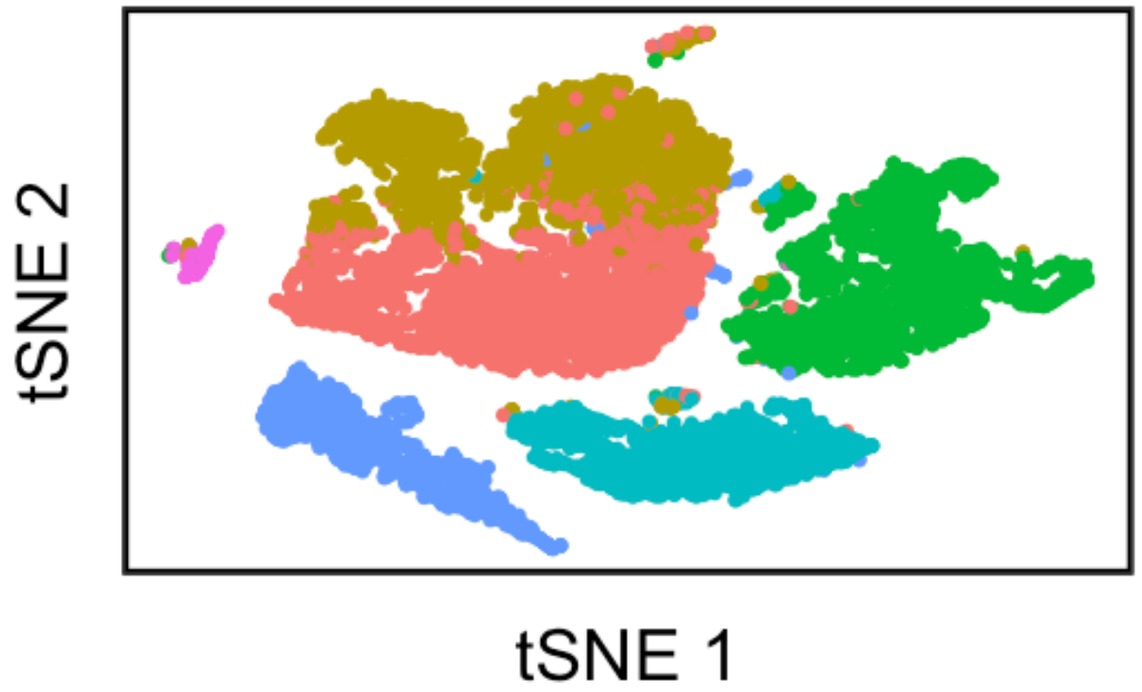


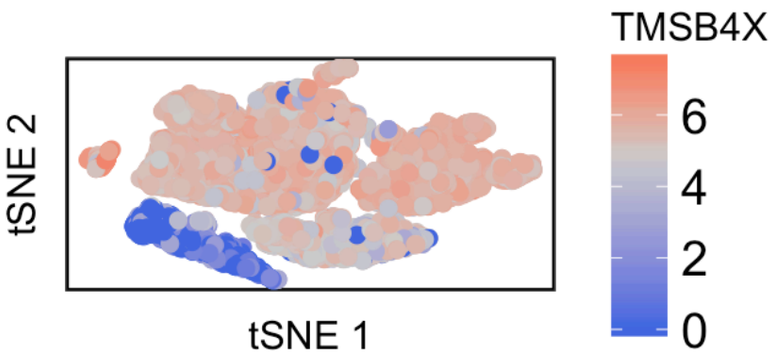


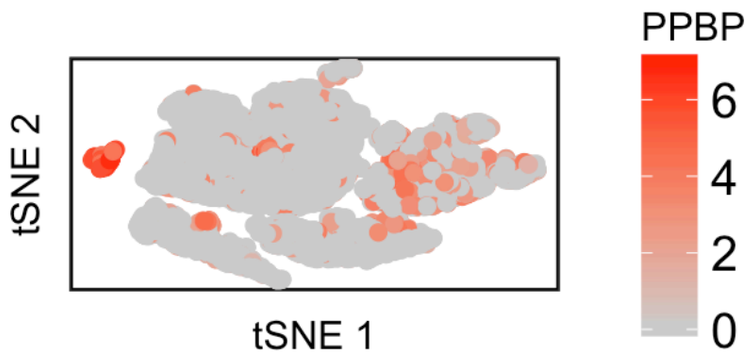


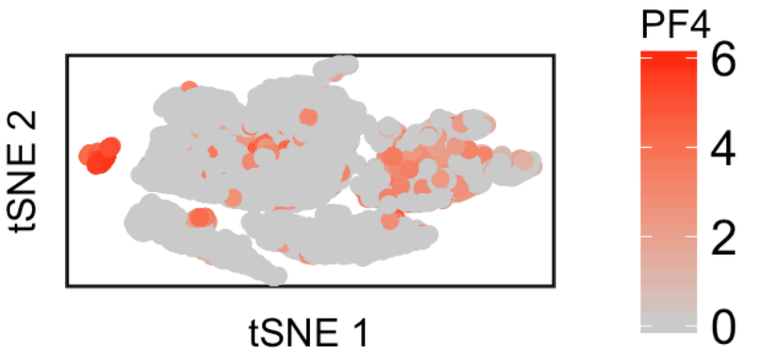
Top enriched pathways

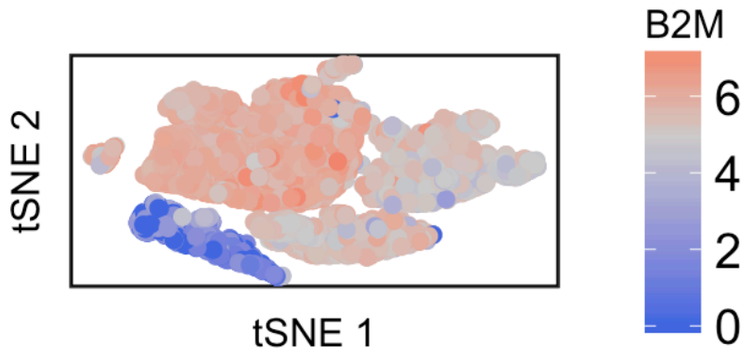
1. hemoglobins chaperone
2. Erythrocytes take up oxygen and release carbon dioxide
3. Erythrocytes take up carbon dioxide and release oxygen
4. O2/CO2 exchange in erythrocytes
5. Malaria - Homo sapiens (human)

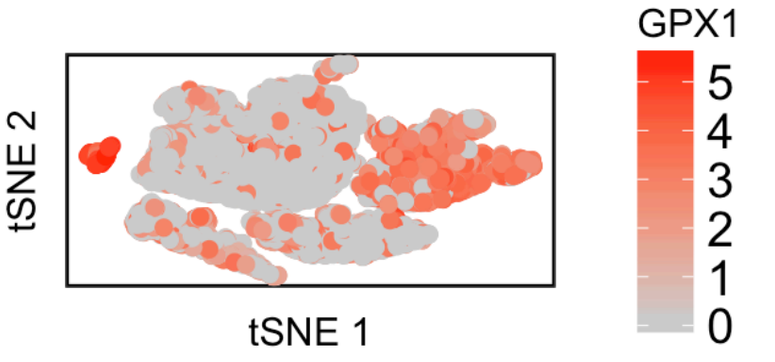


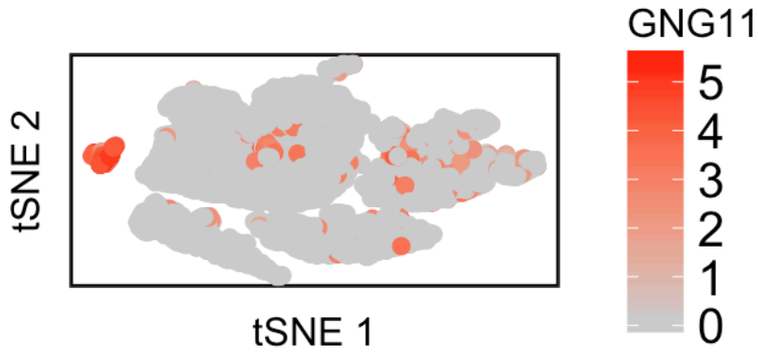








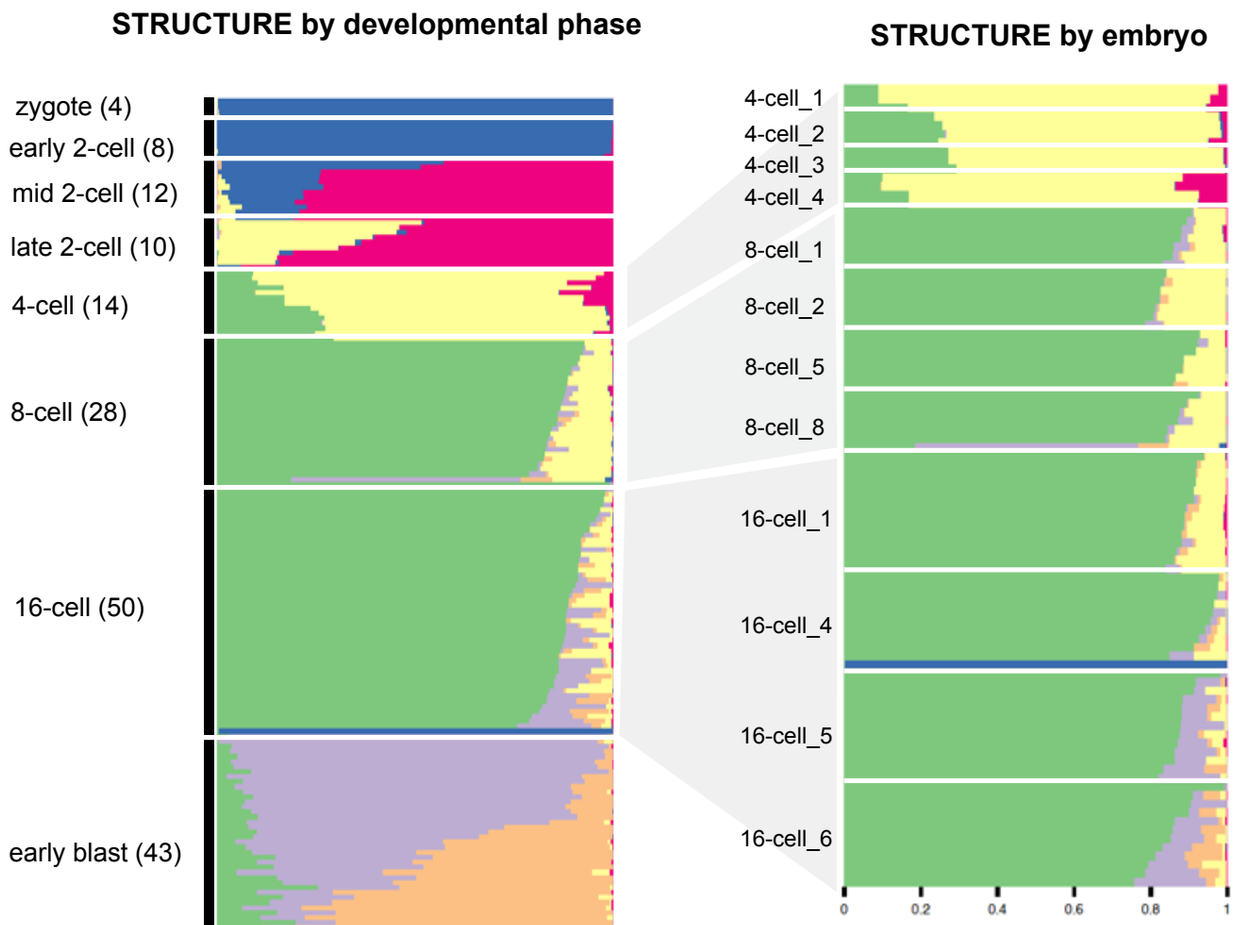




Top enriched pathways

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

Deng et al 2016 single cell developmental study



Clusters	Top 5 driving genes	GO annotations summary
1: Blue	*Bcl2l10, Tcl1 , E330034G19RiK, LOC100502936, Oas1d*	*Gamete generation, Oocyte development & differentiation, Cell devision regulation*
2: Magenta	*Obox3, Zfp352, Gm8300, Usp17l5, BB287469*	*Nuclear body, Centriole, Microtubule organizing center part*
3: Yellow	*Rtn2, Ebna1bp2, Zfp259, Nasp, Cenpe*	*Nuclear part, Nuclear lumen, Intracellular organelle, Nucleolus, Microtubule*
4: Green	*Timd2, Isyna1, Alppl2, Prame15, Hsp90Ab1*	*Cytosol, Cytoplasmic Part, Catabolic Process, Protein Binding, Ubiquitin Ligase*
5: Purple	*Upp1, Tdgf1, Aqp8, Fabp5, Tat*	*Extracellular Exosome/Vesicle/Organelle, Metabolic Process, Blood Vessel Morphogenesis*
6: Orange	*Actb, Krt18, Fabp3, Id2, Tspan8*	*Extracellular Exome/Organelle, Membrane Bound Organelle, Actin Cytoskeleton, Actin Filament Bundle*

Acknowledgements

Matthew Stephens
Chiaowen Joyce Hsiao
Matt Taddy

GTEx Consortium

Raphael Gottardo
Valentin Voillet
Aude Chapuis
Kelly Paulson
Paul Ngheim

Project webpage:

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

http://stephenslab.github.io/count-clustering/

Package : CountClust

devel: https://github.com/kkdey/CountClust