CLUSTERING IN RNA-Seq DATA USING GRADE of MEMBERSHIP MODELS

## Kushal K Dey[1], Chiaowen Joyce Hsiao, Matthew Stephens[1,2]

1   Department of Statistics, University of Chicago, Chicago, Illinois 60637, USA;   2   Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA

Abstract

We use a model based clustering method, similar to "admixture model" in population genetics for interpreting structure in bulk and single cell RNA-seq read counts data.

Each tissue sample in bulk RNA-seq may have memberships in multiple clusters, which are taken as representatives of different cell types.
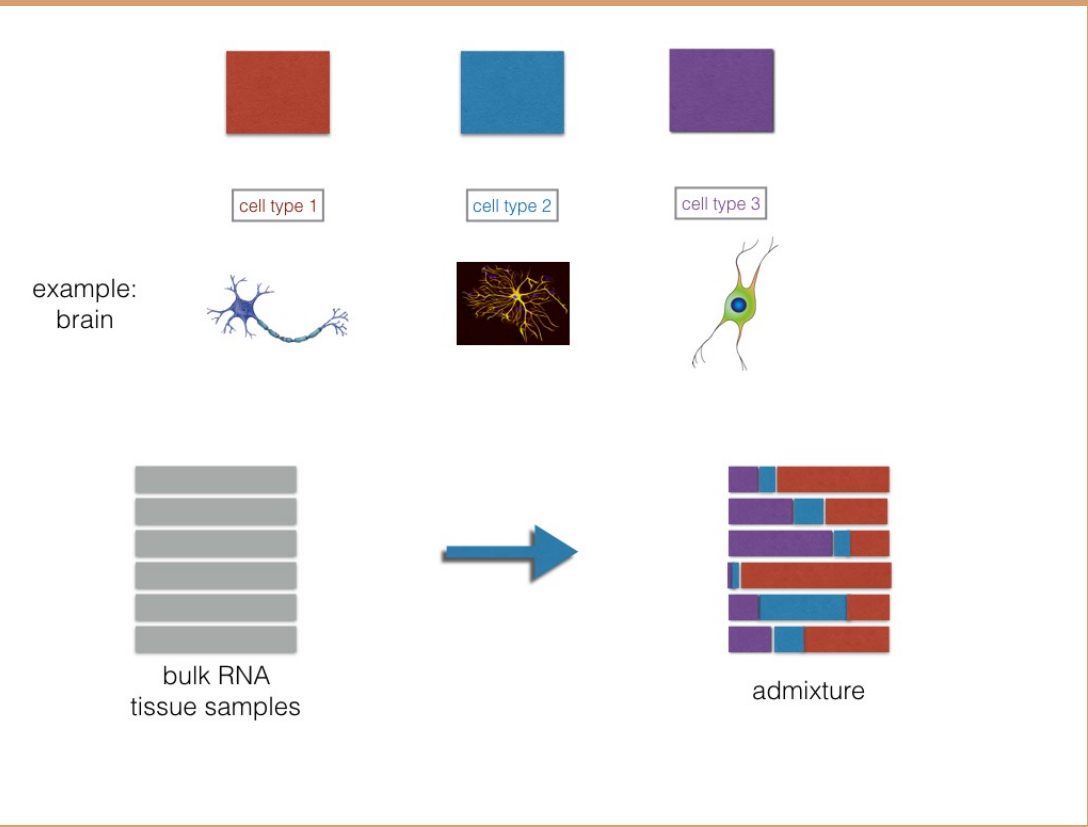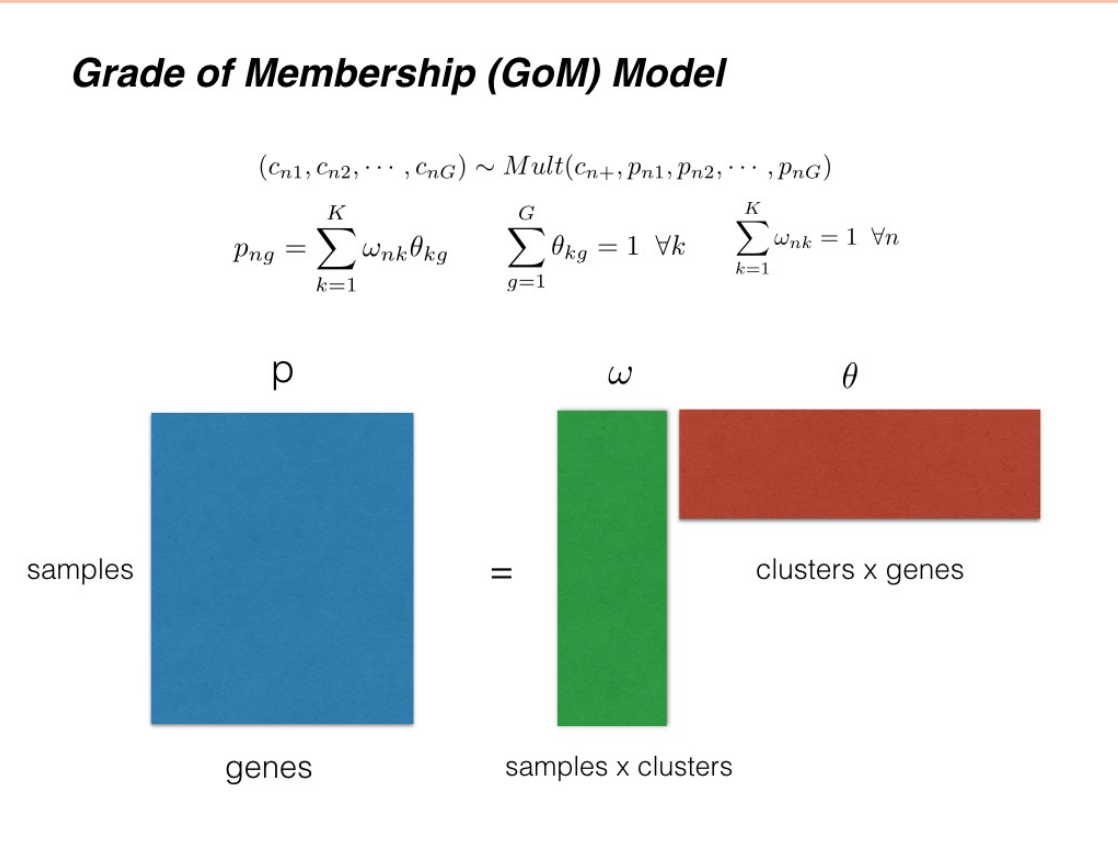
In single cell RNA-seq, each cell may have grades of memberships in the extreme stages of development.
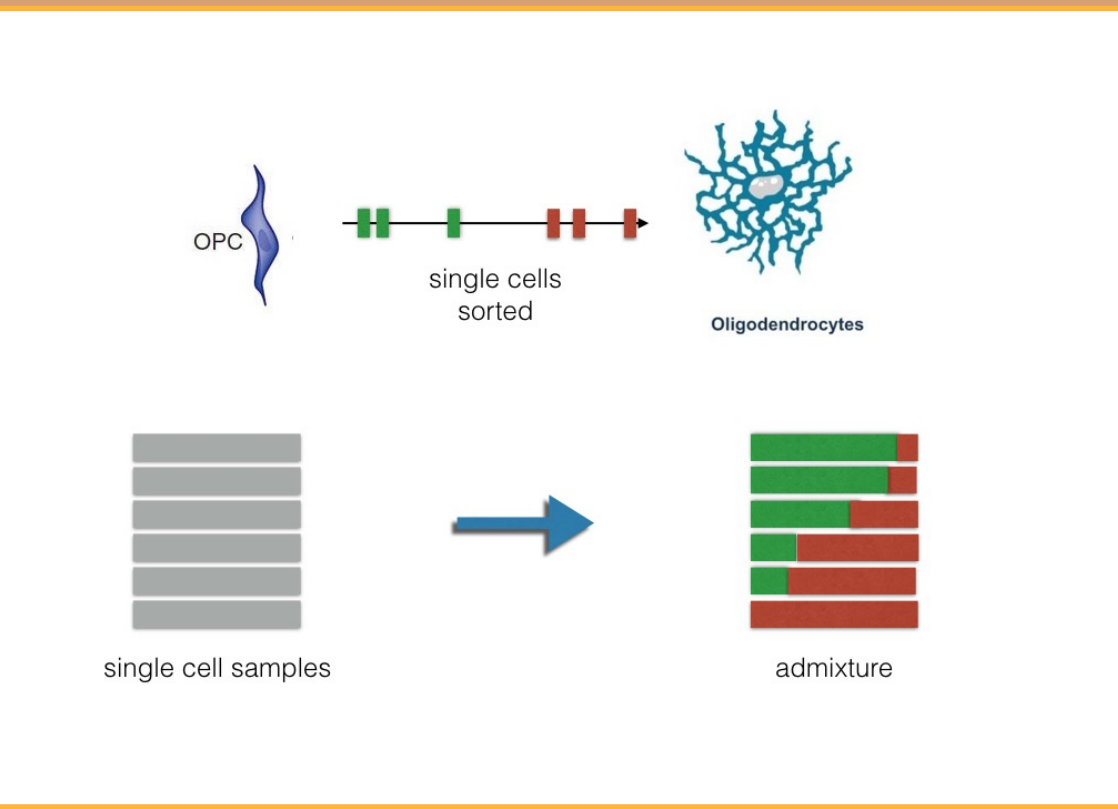
Model

c_{ng} : reads mapped to gene g for sample n
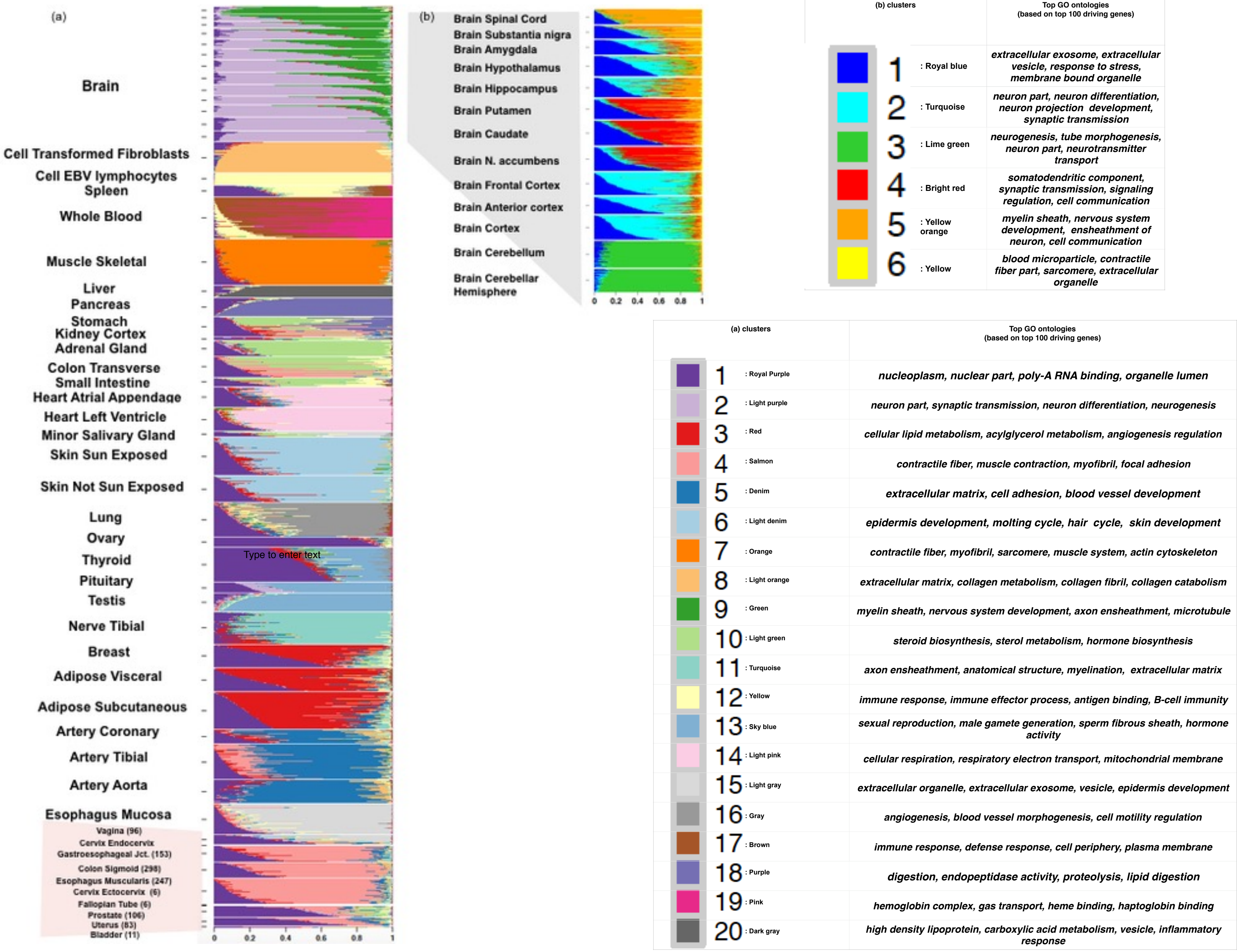
c_{n+} : library size sample n

Grade of Membership (GoM) Model

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

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

p = ω × θ

samples = samples x clusters × clusters x genes
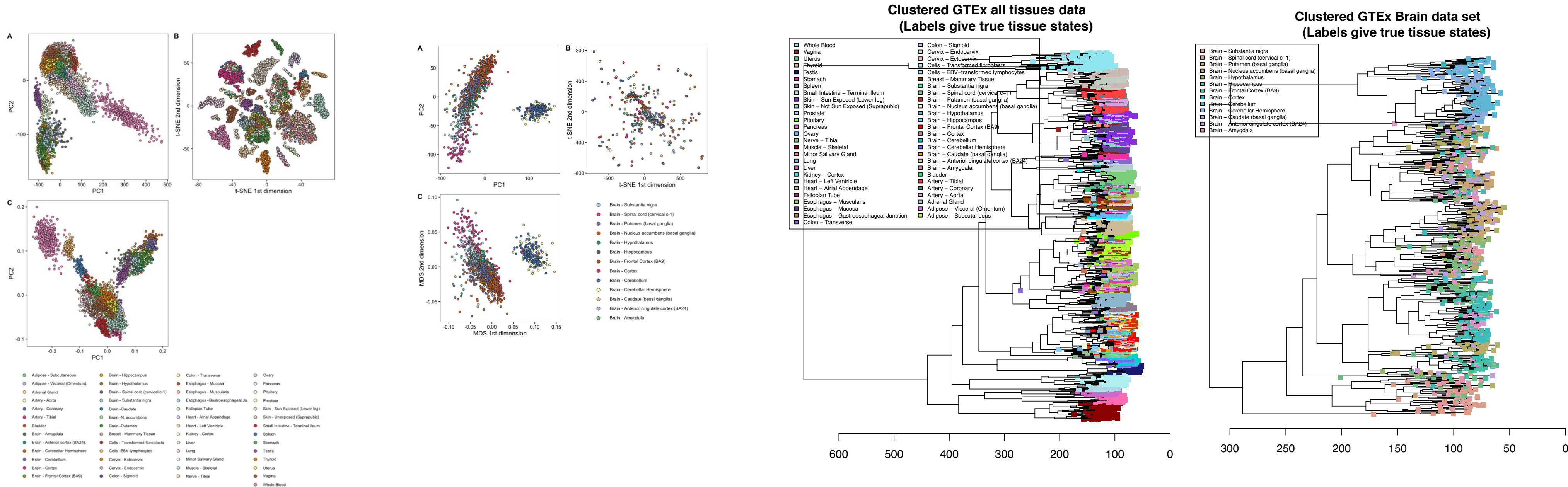
Applications

1. Bulk RNA-seq (GTEx - Genotype tissue expression project bulk RNA samples across tissues)

2. Single cell RNA-seq (Deng et al 2016 - single cell developmental phase data collected from mouse embryo across multiple developmental stages)

The model fit and the visualization using the Structure plot stacked bar chart representation is performed by the **Bioconductor** package **CountClust**
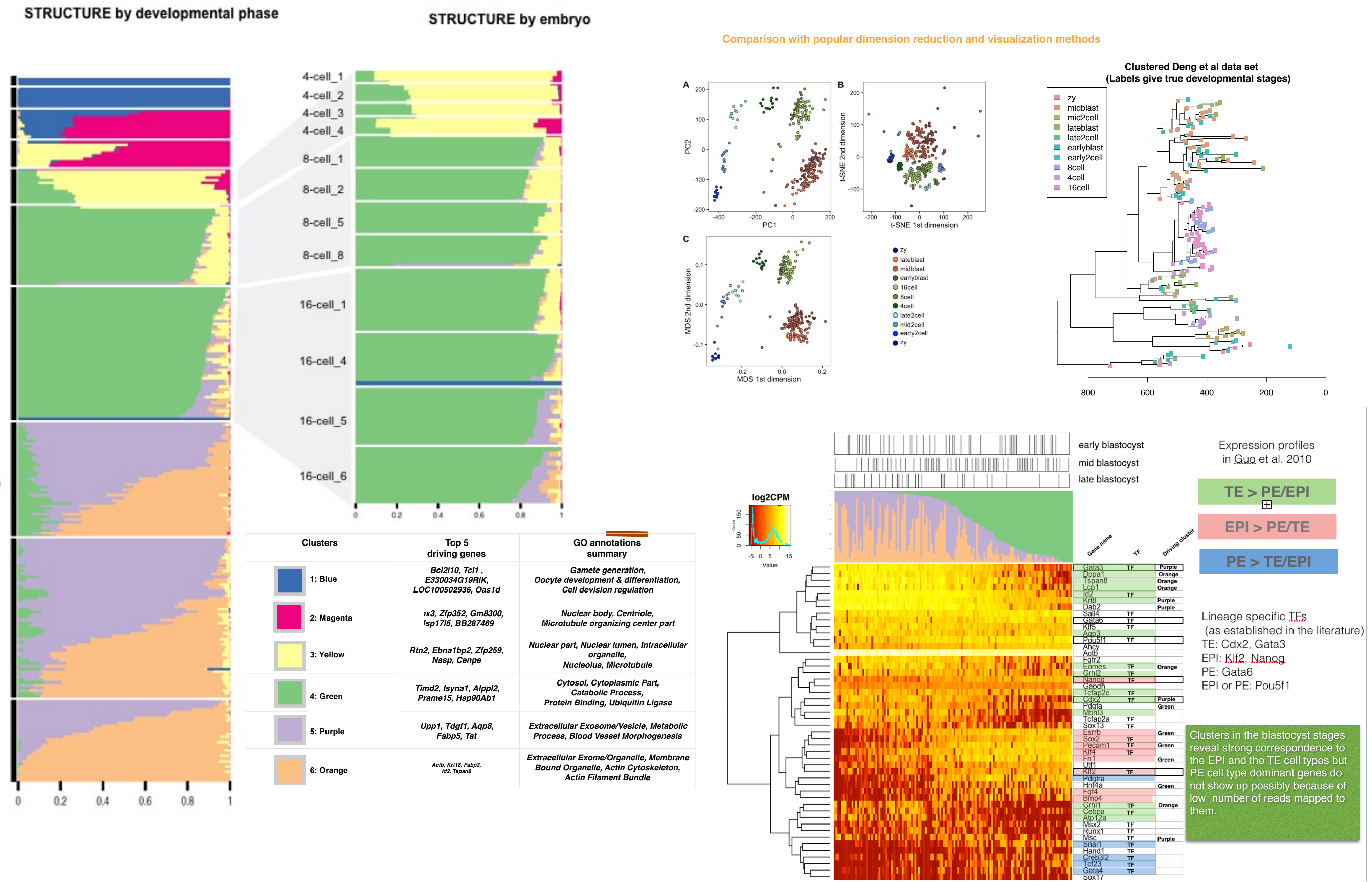
References:
1.   K Dey, C Hsiao, and M Stephens. Clustering RNA-seq expression data using grade of membership model. http://biorxiv.org/content/early/2016/05/03/051631.
2.   Deng Q, Ramskold D, Reinius B, Sandberg R. 2014. Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells. Science. 343 (6167) 193-196.
3.   Guo G, Huss M, Tong GQ, Wang C, Sun LL, Clarke ND, Robson P. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. Developmental Cell. 18(4), 675-685
4.   The GTEx Consortium. 2013. The Genotype-Tissue Expression (GTEx) project. Nature genetics. 45(6): 580-585. doi:10.1038/ng.2653.

Bulk RNA-seq (GTEx data) analysis

(b) clusters — Top GO ontologies (based on top 100 driving genes)

	cluster		Top GO ontologies
	1	: Royal blue	extracellular exosome, extracellular vesicle, response to stress, membrane bound organelle
	2	: Turquoise	neuron part, neuron projection development, synaptic transmission
	3	: Lime green	neurogenesis, tube morphogenesis, neuron part, neurotransmitter transport
	4	: Bright red	somatodendritic component, synaptic transmission, signaling regulation, cell communication
	5	: Yellow orange	myelin sheath, nervous system development, ensheathment of neuron, cell communication
	6	: Yellow	blood microparticle, contractile fiber part, sarcomere, extracellular organelle

(a) clusters — Top GO ontologies (based on top 100 driving genes)

	cluster		Top GO ontologies
	1	: Royal Purple	nucleoplasm, nuclear part, poly-A RNA binding, organelle lumen
	2	: Light purple	neuron part, synaptic transmission, neuron differentiation, neurogenesis
	3	: Red	cellular lipid metabolism, acylglycerol metabolism, angiogenesis regulation
	4	: Salmon	contractile fiber, muscle contraction, myofibril, focal adhesion
	5	: Denim	extracellular matrix, cell adhesion, blood vessel development
	6	: Light denim	epidermis development, molting cycle, hair cycle, skin development
	7	: Orange	contractile fiber, myofibril, sarcomere, muscle system, actin cytoskeleton
	8	: Light orange	extracellular matrix, collagen metabolism, collagen fibril, collagen catabolism
	9	: Green	myelin sheath, nervous system development, axon ensheathment, microtubule
	10	: Light green	steroid biosynthesis, sterol metabolism, hormone biosynthesis
	11	: Turquoise	axon ensheathment, anatomical structure, myelination, extracellular matrix
	12	: Yellow	immune response, immune effector process, antigen binding, B-cell immunity
	13	: Sky blue	sexual reproduction, male gamete generation, sperm fibrous sheath, hormone activity
	14	: Light pink	cellular respiration, respiratory electron transport, mitochondrial membrane
	15	: Light gray	extracellular organelle, extracellular exosome, vesicle, epidermis development
	16	: Gray	angiogenesis, blood vessel morphogenesis, cell motility regulation
	17	: Brown	immune response, defense response, cell periphery, plasma membrane
	18	: Purple	digestion, endopeptidase activity, proteolysis, lipid digestion
	19	: Pink	hemoglobin complex, gas transport, heme binding, haptoglobin binding
	20	: Dark gray	high density lipoprotein, carboxylic acid metabolism, vesicle, inflammatory response

Comparison with popular dimension reduction and visualization methods

Clustered GTEx all tissues data (Labels give true tissue states)

Clustered GTEx Brain data set (Labels give true tissue states)

Single cell RNA-seq analysis (Deng et al 2016)

STRUCTURE by developmental phase

STRUCTURE by embryo

Comparison with popular dimension reduction and visualization methods

Clustered Deng et al data set (Labels give true developmental stages)

Clusters	Top 5 driving genes	GO annotations summary
1: Blue	Bcl2l10, Tcl1 , E330034G19Rik, LOC100502936, Oas1d	Gamete generation, Oocyte development & differentiation, Cell division regulation
2: Magenta	x2, Zfp352, Gm8300, *sp1715, BB287469	Nuclear body, Centriole, Microtubule organizing center part
3: Yellow	Rtn2, Edna1bp2, 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, Metabolic Process, Blood Vessel Morphogenesis
6: Orange	Actb, Krt18, Fabp3, Id2, Tspan8	Extracellular Exosome/Organelle, Membrane Bound Organelle, Actin Cytoskeleton, Actin Filament Bundle

log2CPM

early blastocyst
mid blastocyst
late blastocyst

Expression profiles in Guo et al. 2010

TE > PE/EPI

EPI > PE/TE

PE > TE/EPI

Lineage specific TFs (as established in the literature)
TE: Cdx2, Gata3
EPI: Klf2, Nanog
PE: Gata6
EPI or PE: Pou5f1

Clusters in the blastocyst stages reveal strong correspondence to the EPI and the TE cell types but PE cell type dominant genes do not show up possibly because of low number of reads mapped to them.