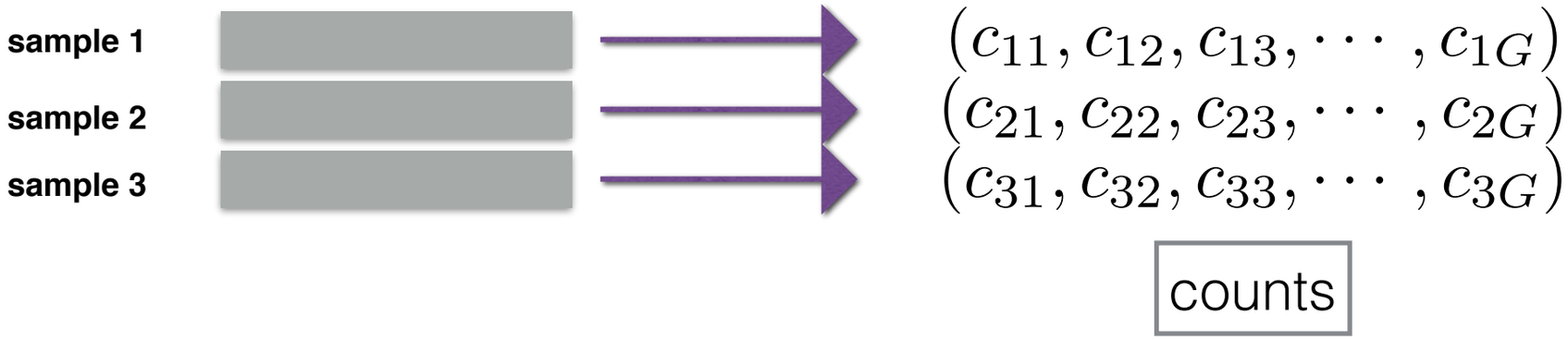
Model-Based Clustering and Visualization of RNA-Seq Data

Kushal Dey
Joyce Hsiao
Matthew Stephens

RNA-seq read counts data



- We consider number of read counts per gene
- Samples may come from bulk RNA-seq (tissue samples) or single cell RNA-seq (scRNA-seq).

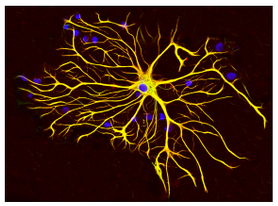
Bulk RNA-seq

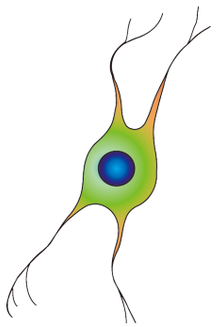
cell type 1

cell type 2

cell type 3

example:
brain





Neuron

Astrocyte

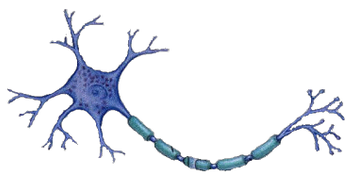
Oligodendrocytes

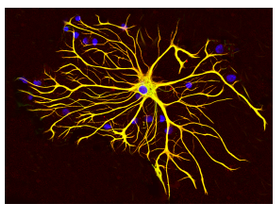
cell type 1

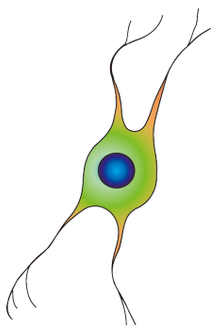
cell type 2

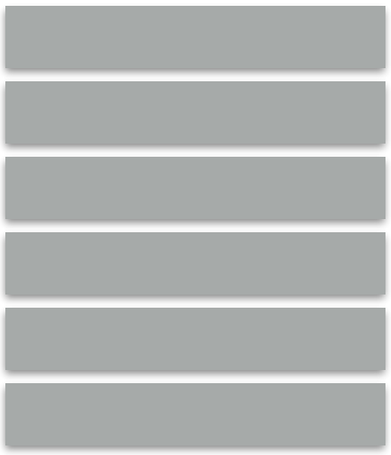
cell type 3

example:
brain

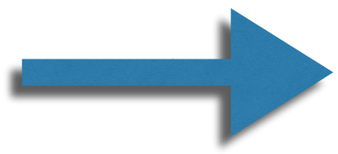


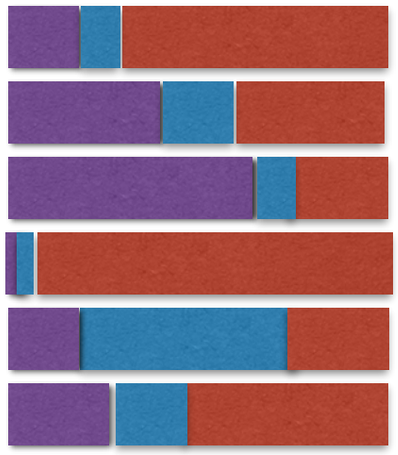






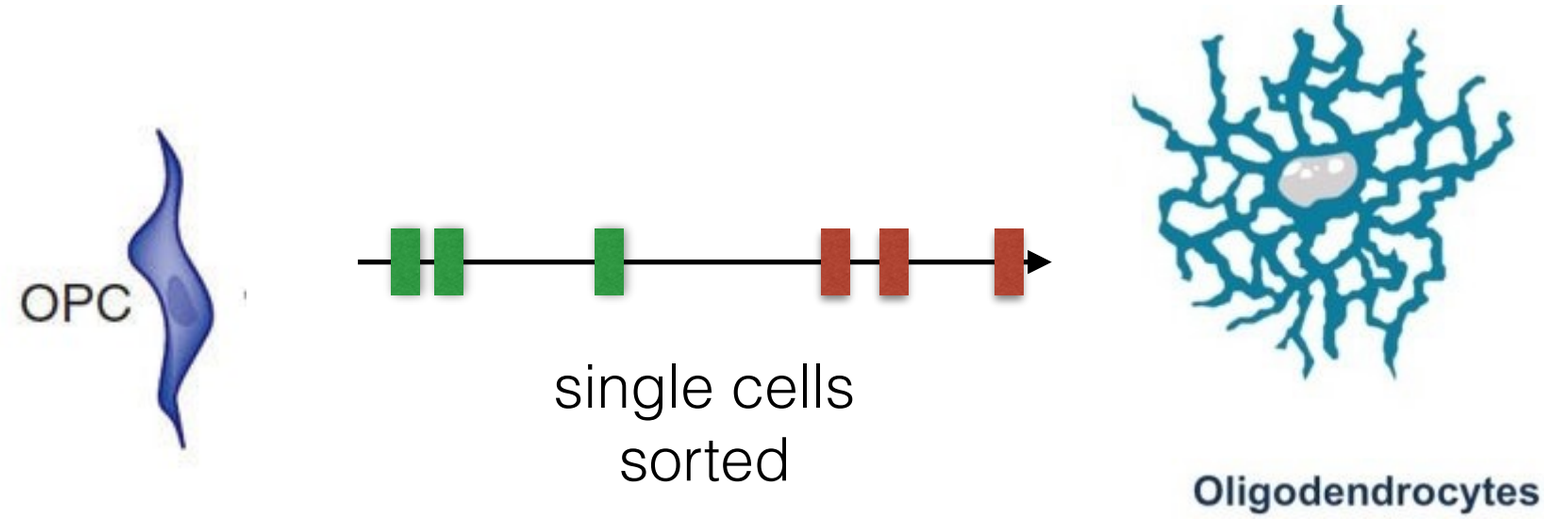
bulk RNA
tissue samples

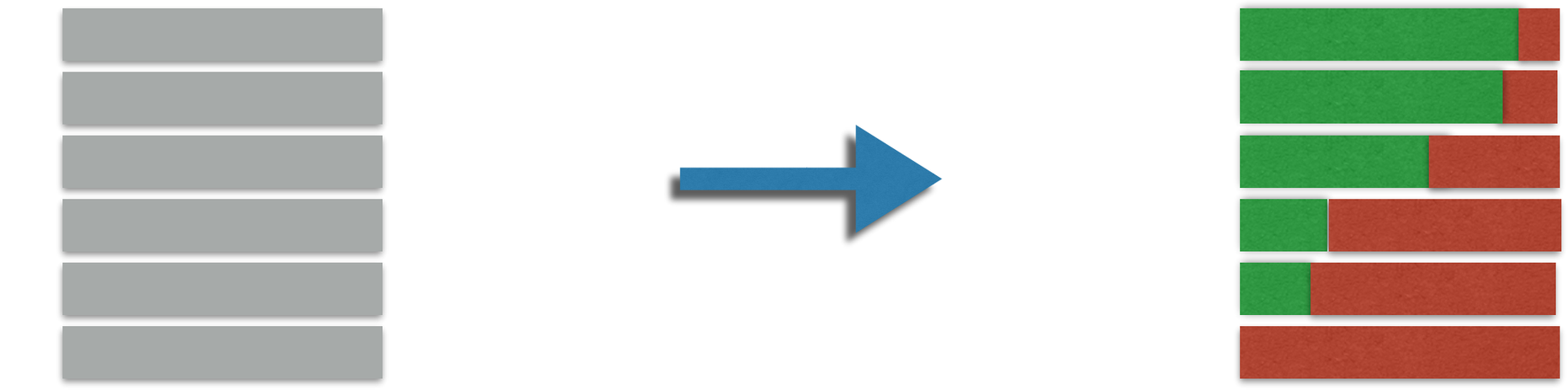




admixture

Single cell RNA-seq





single cell samples

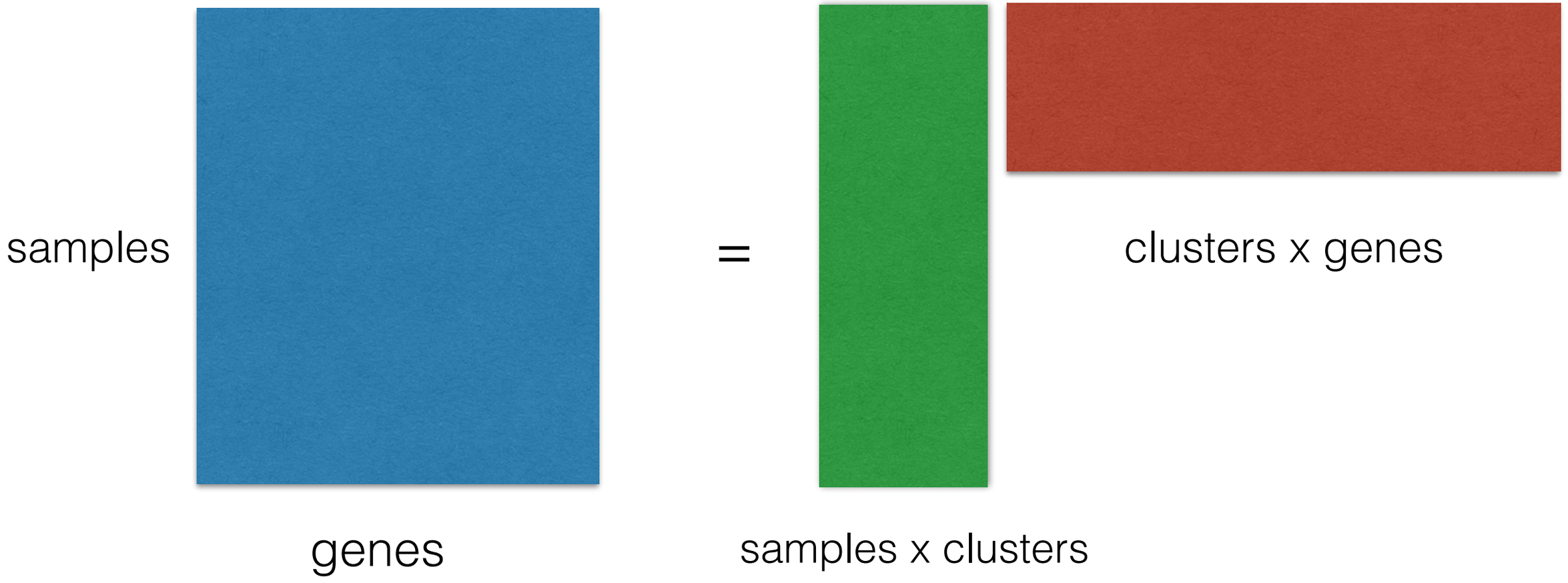
admixture

Model

Grade of Membership (GoM) Model

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

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



Grade of Membership Model (continued)

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

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

ω_{nk} grade of membership of sample n in cluster or topic k

θ_{kg} relative proportion of expression of gene g in cluster k

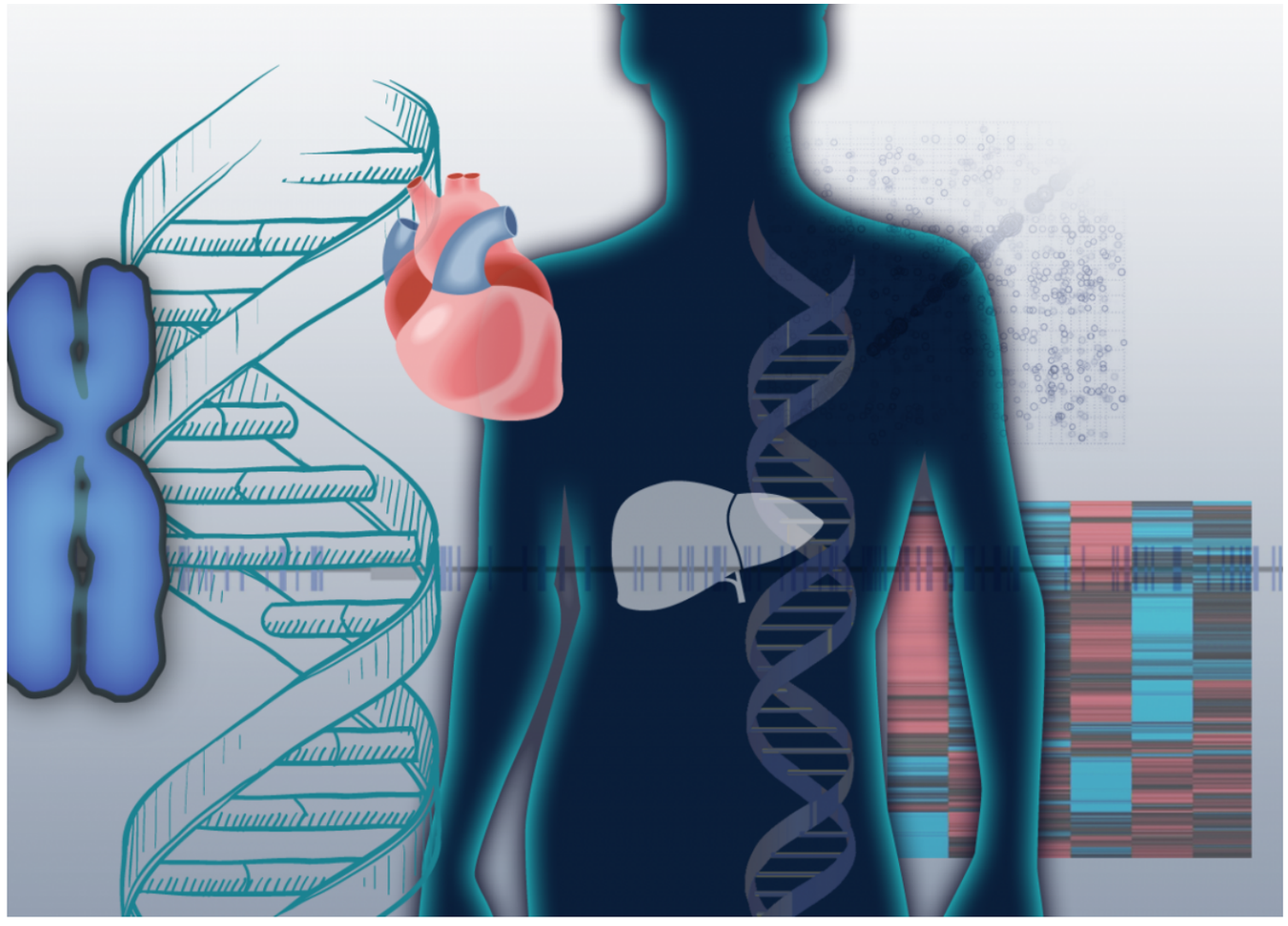
ω_{nk} and θ_{kg} were determined in an unsupervised manner assuming Dirichlet priors

$$\omega_{n*} \sim Dir\left(\frac{1}{K}, \frac{1}{K}, \cdots, \frac{1}{K}\right) \qquad \theta_{k*} \sim Dir\left(\frac{1}{KG}, \frac{1}{KG}, \cdots, \frac{1}{KG}\right)$$

Data Analysis

Genotype Tissue Expression Project - V6

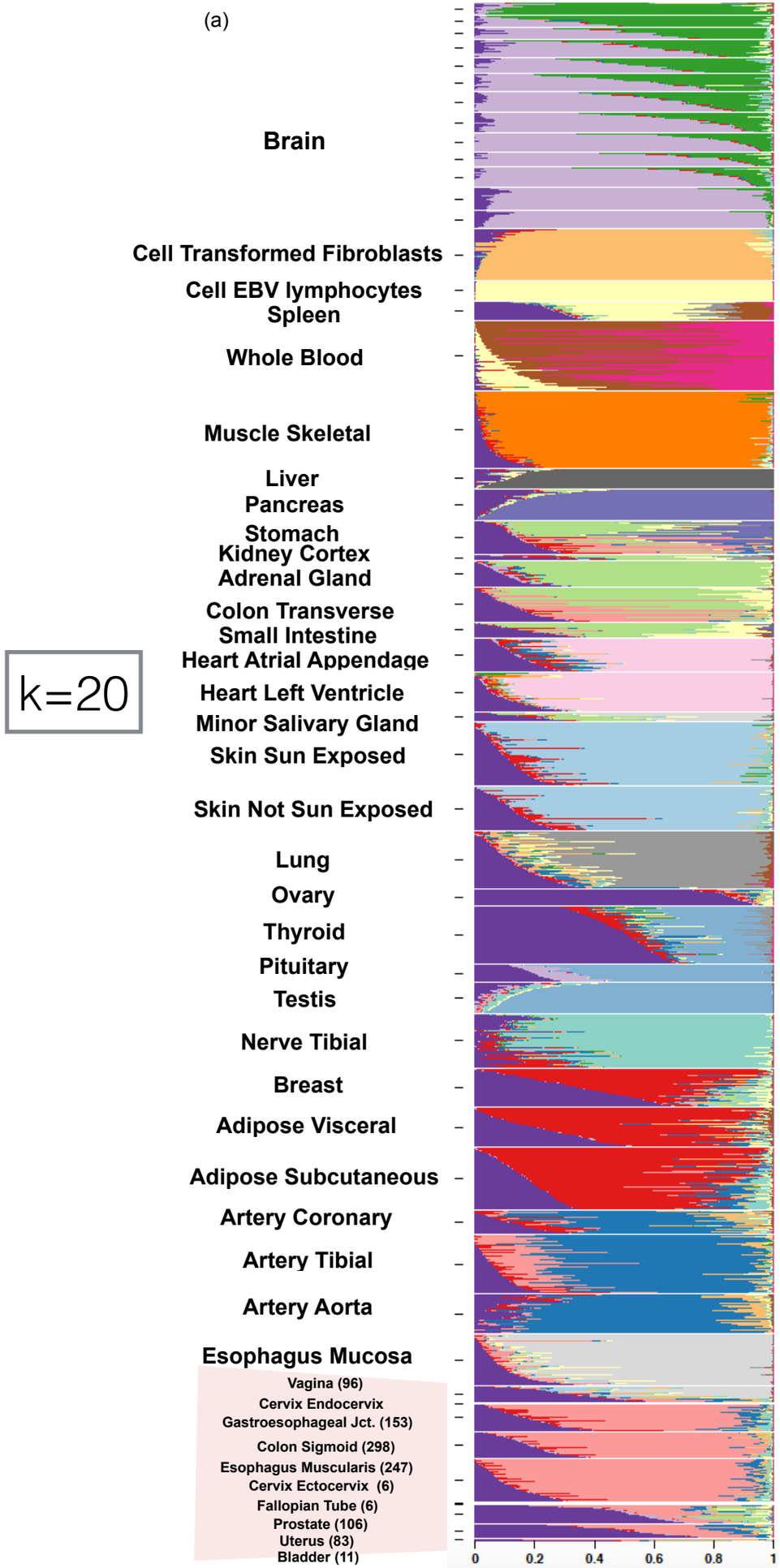
has data on 8555 tissue samples taken from 900 donors (Phase 6) for around 56,000 genes.



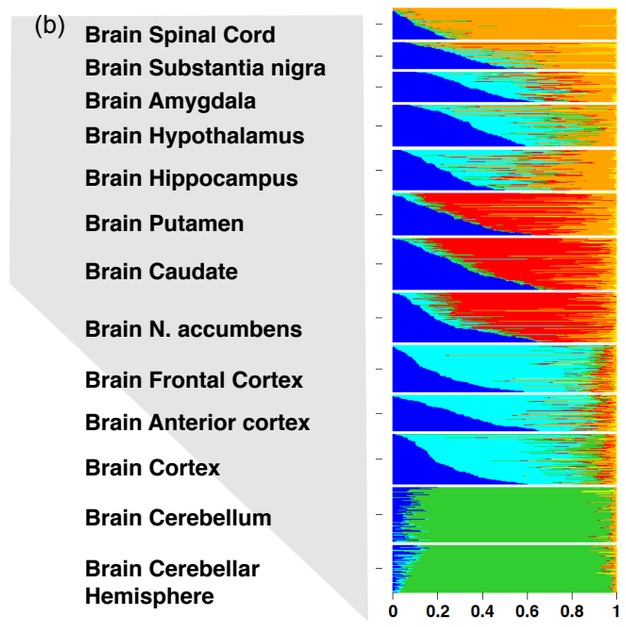
The tissue samples came from a total of 53 different tissues/ sub-tissues (e.g. - brain sub tissues, whole blood etc)

We filtered out around 16,069 genes that passed QC tests.

We fit the GoM model and get estimates of ω and θ



k=20



k=6

Each row is a plot of the vector

ω_{n*} for each sample n

The colors correspond to different clusters

The results presented is corresponding to the GoM fit with highest Bayes factor out of a total of 5 runs.

Extracting top genes per cluster

We compute for each cluster k and each gene g

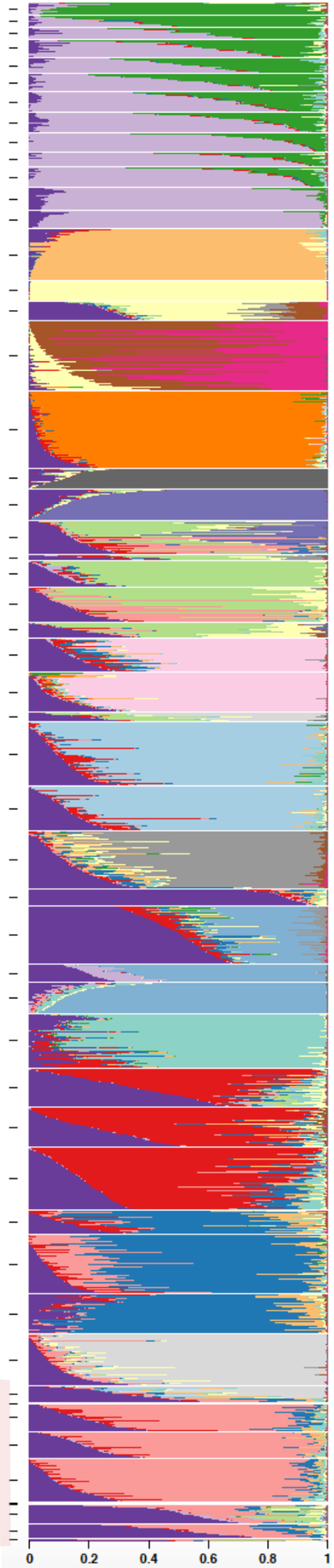
$$D^g[k] = \min_{l \neq k} \left[\hat{\theta}_{kg} log \left(\frac{\hat{\theta}_{kg}}{\hat{\theta}_{lg}} \right) + \hat{\theta}_{lg} - \hat{\theta}_{kg} \right]$$

For each k, choose the genes with the highest values of $D^g[k]$

which satisfy $\hat{\theta}_{kg} = \max_l \ \hat{\theta}_{lg}$

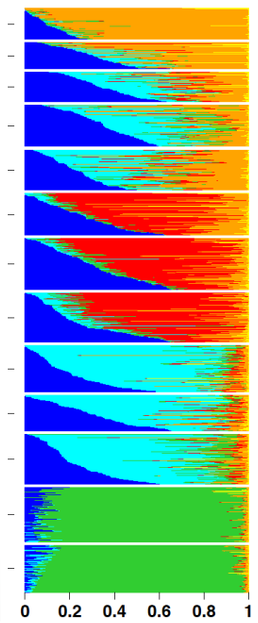
(a)

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)



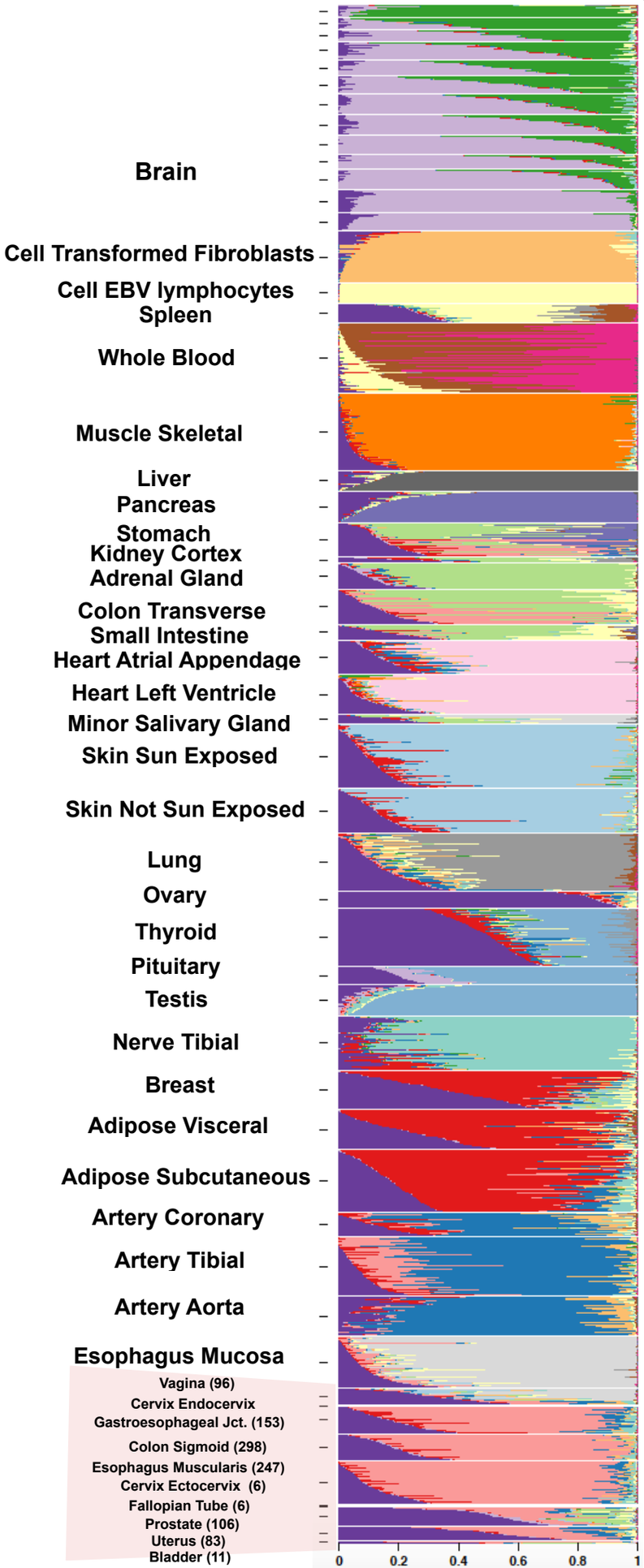
(b)

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

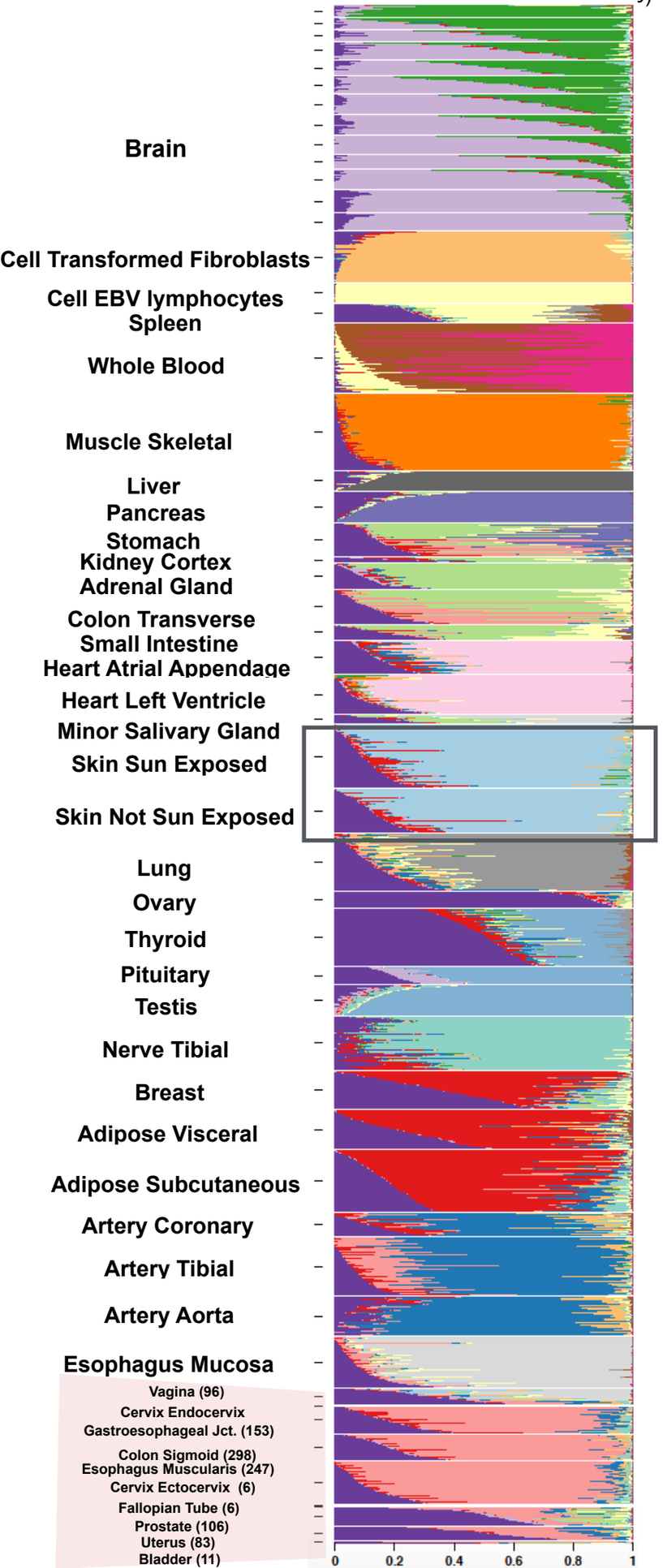


(a) clusters	Top 5 driving genes
1 : Royal Purple	*NEAT1, IGFBP5, CCLN2, SRSF5, PNISR*
2 : Light purple	*SNAP25, FBXL16, NCDN, SNCB, SLC17A7*
3 : Red	*FABP4, PLIN1, FASN, GPX3, LIPE*
4 : Salmon	*ACTG2, MYH11, SYNM, MYLK, CSRP1*
5 : Denim	*RGS5, MGP, AEBP1, IGFBP7, MFGE5*
6 : Light denim	*KRT10, KRT1, KRT2, LOR, KRT14*
7 : Orange	*NEB, MYH1, MYH2, MYBPC1, ACTA1*
8 : Light orange	*FN1, COL1A1, COL1A2, COL3A1, COL6A3*
9 : Green	*MBP, GFAP, MTURN, H1PK2, CARNS1*
10 : Light green	*CYP17A1, CYP11B1, PIGR, GKN1, STAR*
11 : Turquoise	*MPZ, APOD, PMP22, PRX, NGFR*
12 : Yellow	*IGHM, IGHG1, IGHG2, IGHG4, CD74*
13 : Sky blue	*TG, PRL, GH1, PRM2, PRM1*
14 : Light pink	*NPPA, MYH6, TNNT2, ACTC1, MYBPC3*
15 : Light gray	*KRT13, KRT4, MUC7, CRNN, KRT6A*
16 : Gray	*SFTPB, SFTPA1, SFTPA2, SFTPC, A2M*
17 : Brown	*CSF3R, MMP25, IL1R2, SELL, VNN2*
18 : Purple	*PRSS1, CPA1, PNLIP, CELA3A, GP2*
19 : Pink	*HBB, HBA2, HBA1, FKBP8, HBD*
20 : Dark gray	*ALB, HP, FGB, FGA, ORM1*

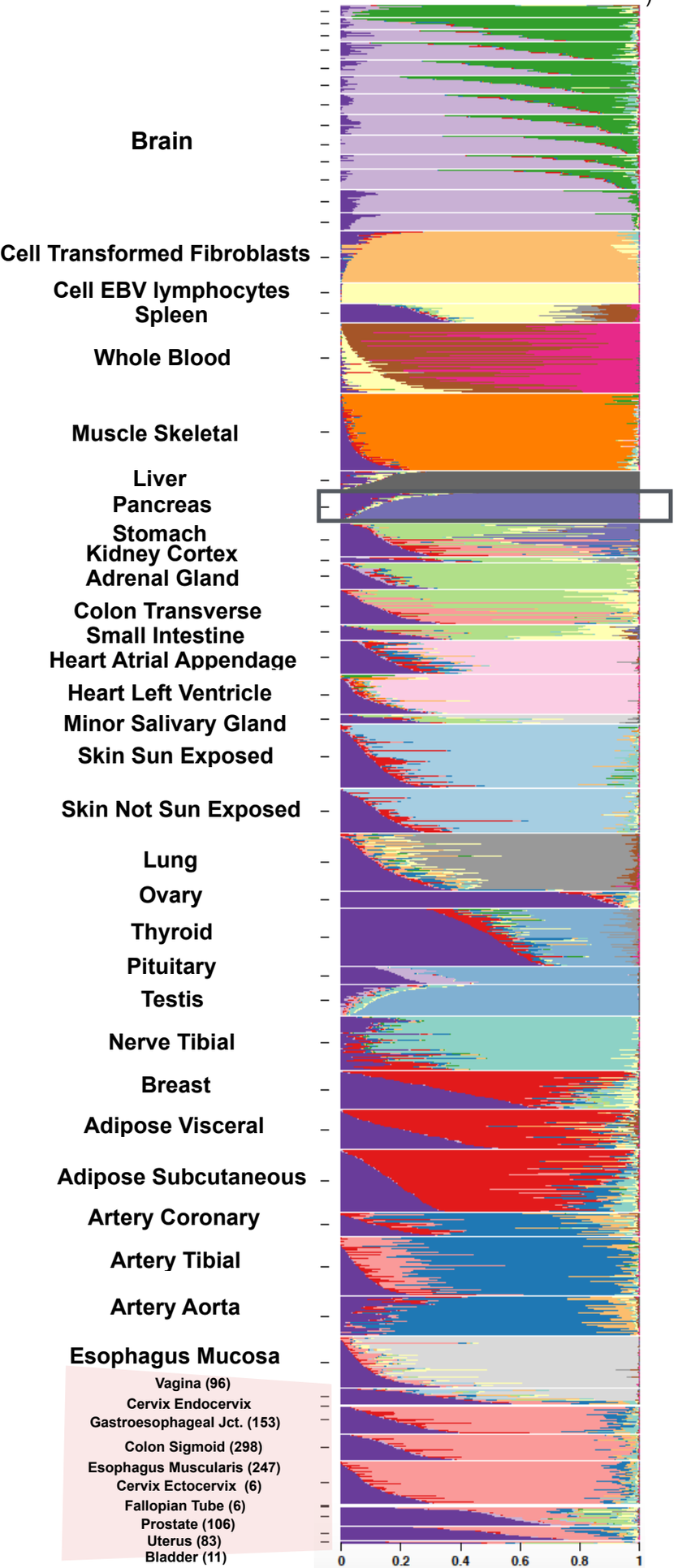
(b) clusters	Top 5 driving genes
1 : Royal blue	*CLU, OXT, GLUL, NDRG2, CST3*
2 : Turquoise	*ENC1, NCALD, YWHAH, KIF5A, NPTXR*
3 : Lime green	*PKD1, CBLN3, CHGB, COL27A1, ABLIM1*
4 : Bright red	*PPP1R1B, RGS14, NCDN, PDE1B, RAP1GAP*
5 : Yellow orange	*MBP, GFAP, TF, MTURN, SCD..*
6 : Yellow	*IQGAP1, A2M, C3, MYHH7, TG*



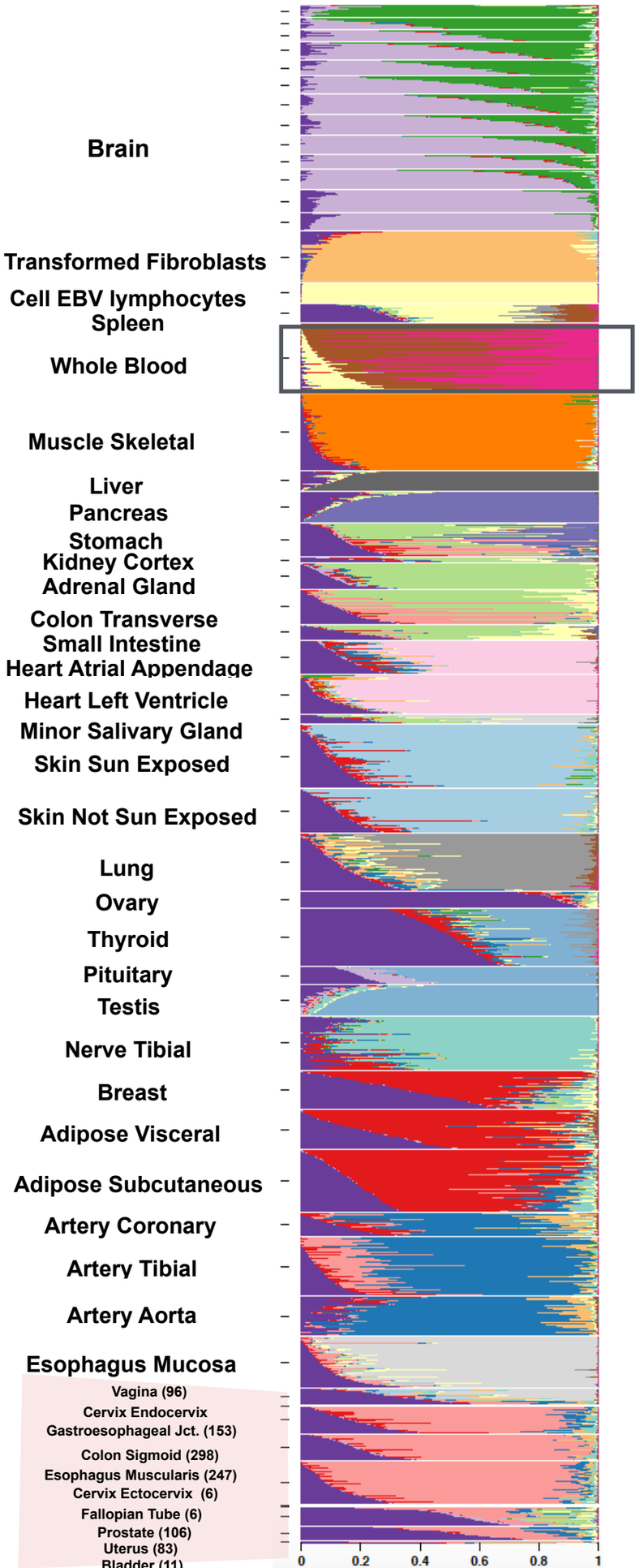
(a) clusters	Top GO ontologies (based on top 100 driving genes)
1 : Royal Purple	*nucleoplasm, nuclear part, poly-A RNA binding, organelle lumen*
2 : Light purple	*neuron part, synaptic transmission, neuron differentiation, neurogenesis*
3 : Red	*cellular lipid metabolism, acylglycerol metabolism, angiogenesis 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*



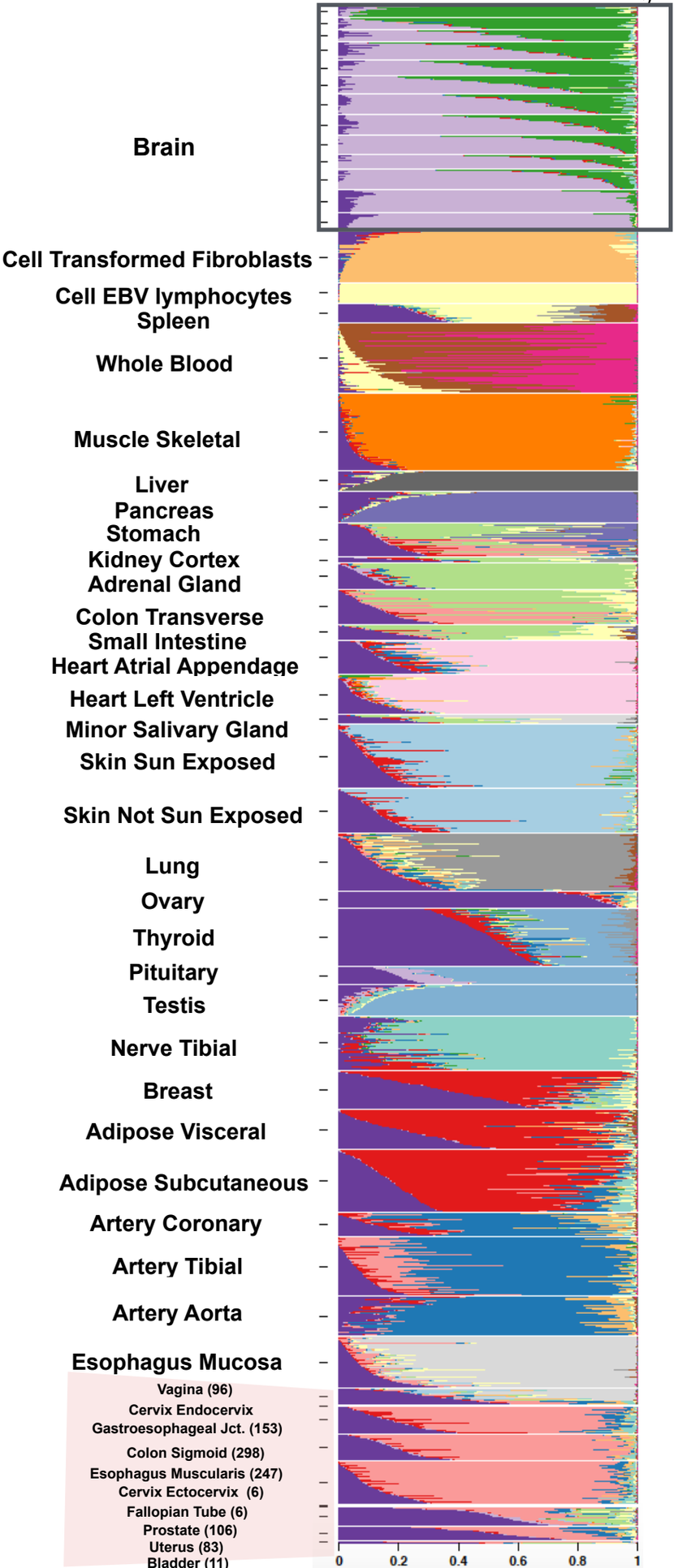
(a) clusters	Top GO ontologies (based on top 100 driving genes)
1 : Royal Purple	*nucleoplasm, nuclear part, poly-A RNA binding, organelle lumen*
2 : Light purple	*neuron part, synaptic transmission, neuron differentiation, neurogenesis*
3 : Red	*cellular lipid metabolism, acylglycerol metabolism, angiogenesis 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*



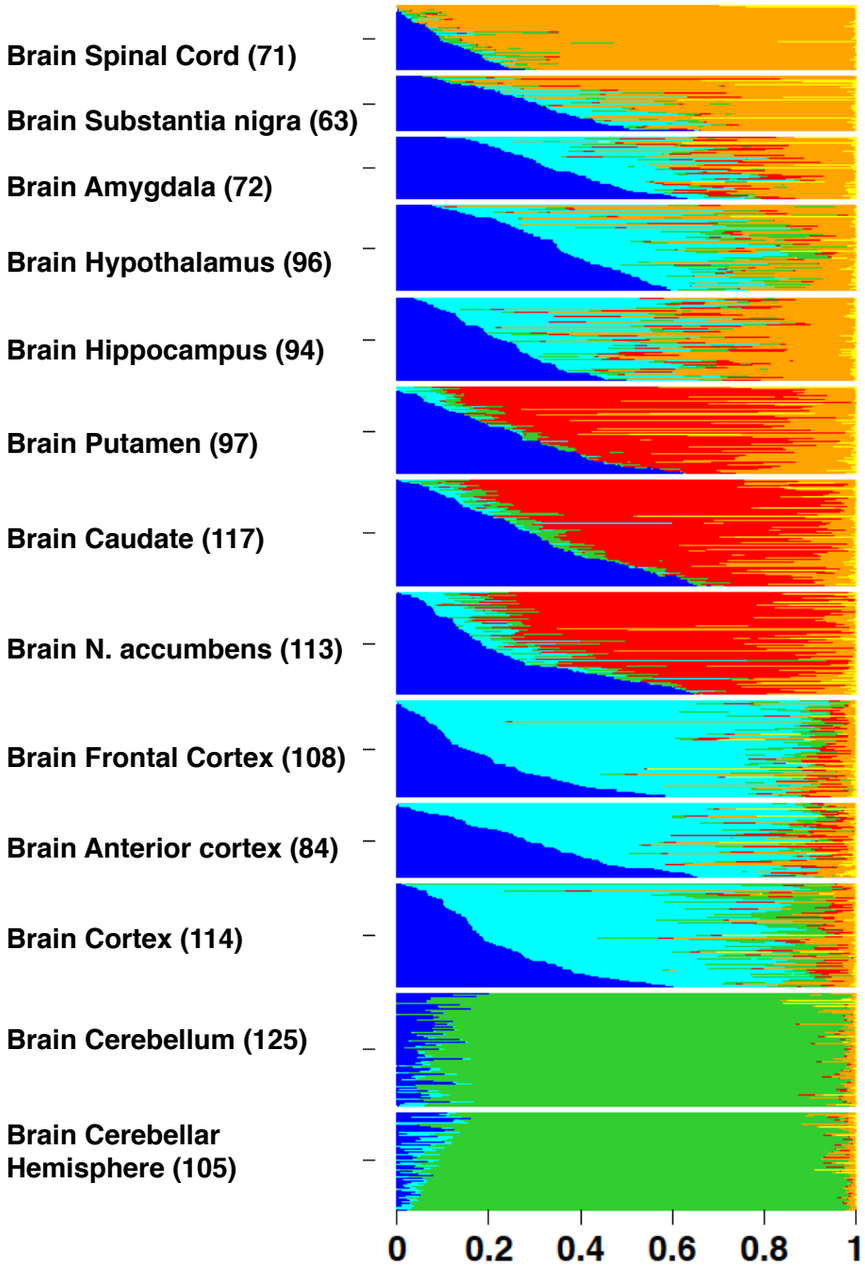
(a) clusters	Top GO ontologies (based on top 100 driving genes)
1 : Royal Purple	*nucleoplasm, nuclear part, poly-A RNA binding, organelle lumen*
2 : Light purple	*neuron part, synaptic transmission, neuron differentiation, neurogenesis*
3 : Red	*cellular lipid metabolism, acylglycerol metabolism, angiogenesis 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*



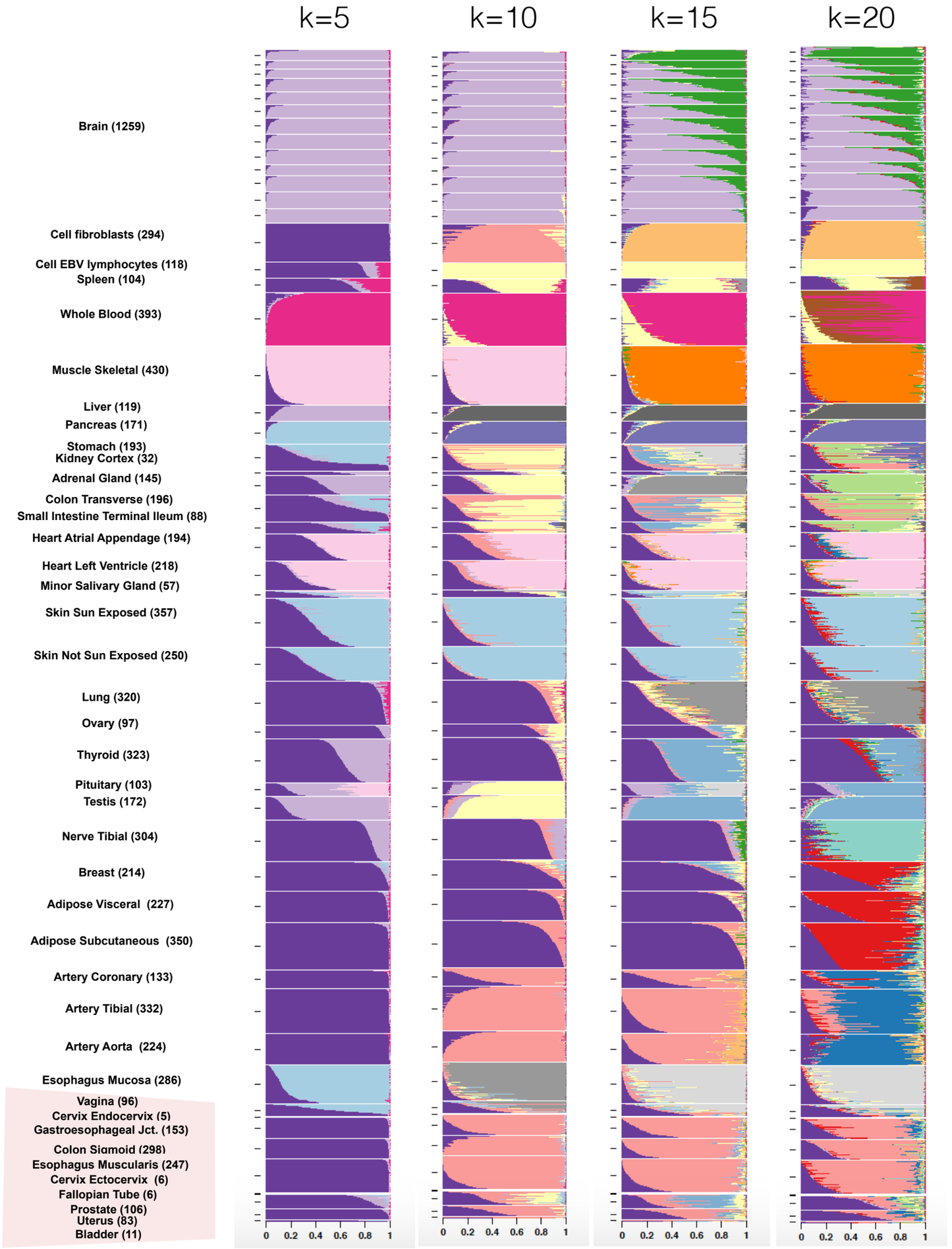
(a) clusters	Top GO ontologies (based on top 100 driving genes)
1 : Royal Purple	*nucleoplasm, nuclear part, poly-A RNA binding, organelle lumen*
2 : Light purple	*neuron part, synaptic transmission, neuron differentiation, neurogenesis*
3 : Red	*cellular lipid metabolism, acylglycerol metabolism, angiogenesis 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*



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



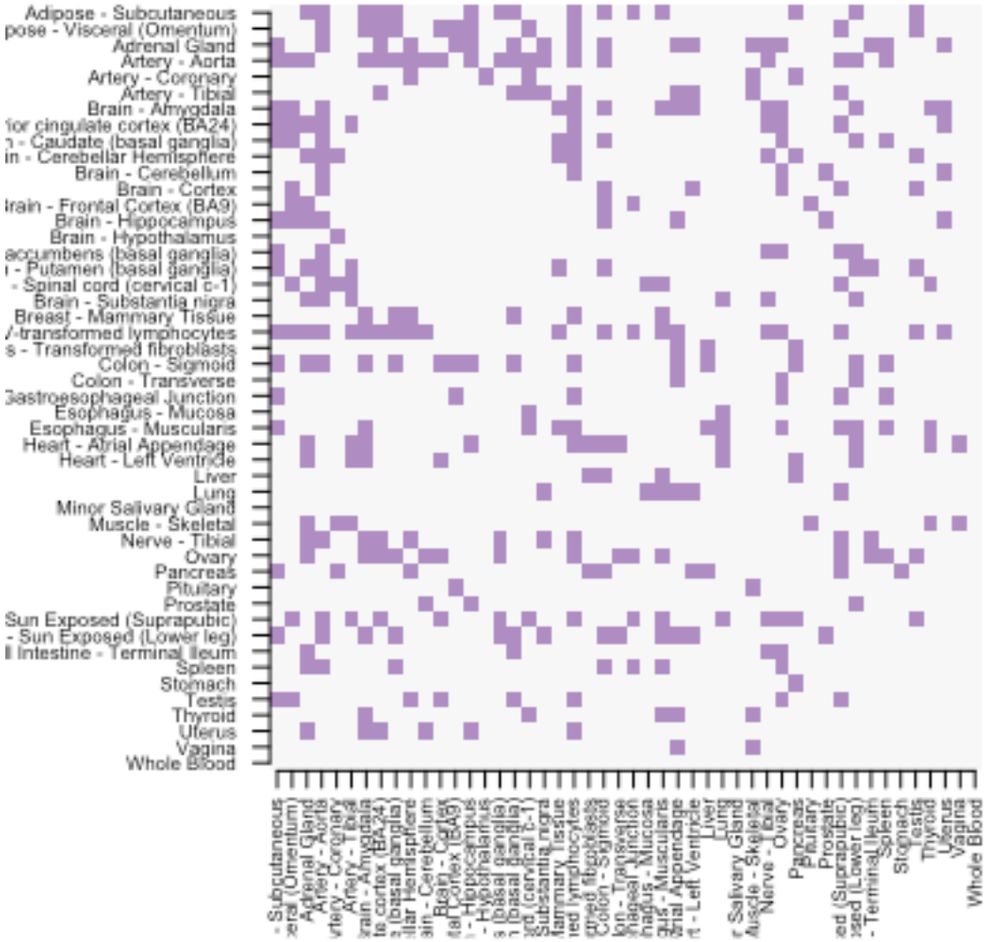
(b) clusters		Top GO ontologies (based on top 100 driving genes)
1	: Royal blue	***extracellular exosome, extracellular vesicle, response to stress, membrane bound organelle***
2	: Turquoise	***neuron part, neuron differentiation, neuron projection development, synaptic transmission***
3	: Lime green	***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 regulation***
6	: Yellow	***blood microparticle, contractile fiber part, sarcomere, extracellular organelle***


k=5
k=10
k=15
k=20
Brain (1259)
Cell fibroblasts (294)
Cell EBV lymphocytes (118)
Spleen (104)
Whole Blood (393)
Muscle Skeletal (430)
Liver (119)
Pancreas (171)
Stomach (193)
Kidney Cortex (32)
Adrenal Gland (145)
Colon Transverse (196)
Small Intestine Terminal Ileum (88)
Heart Atrial Appendage (194)
Heart Left Ventricle (218)
Minor Salivary Gland (57)
Skin Sun Exposed (357)
Skin Not Sun Exposed (250)
Lung (320)
Ovary (97)
Thyroid (323)
Pituitary (103)
Testis (172)
Nerve Tibial (304)
Breast (214)
Adipose Visceral (227)
Adipose Subcutaneous (350)
Artery Coronary (133)
Artery Tibial (332)
Artery Aorta (224)
Esophagus Mucosa (286)
Vagina (96)
Cervix Endocervix (5)
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

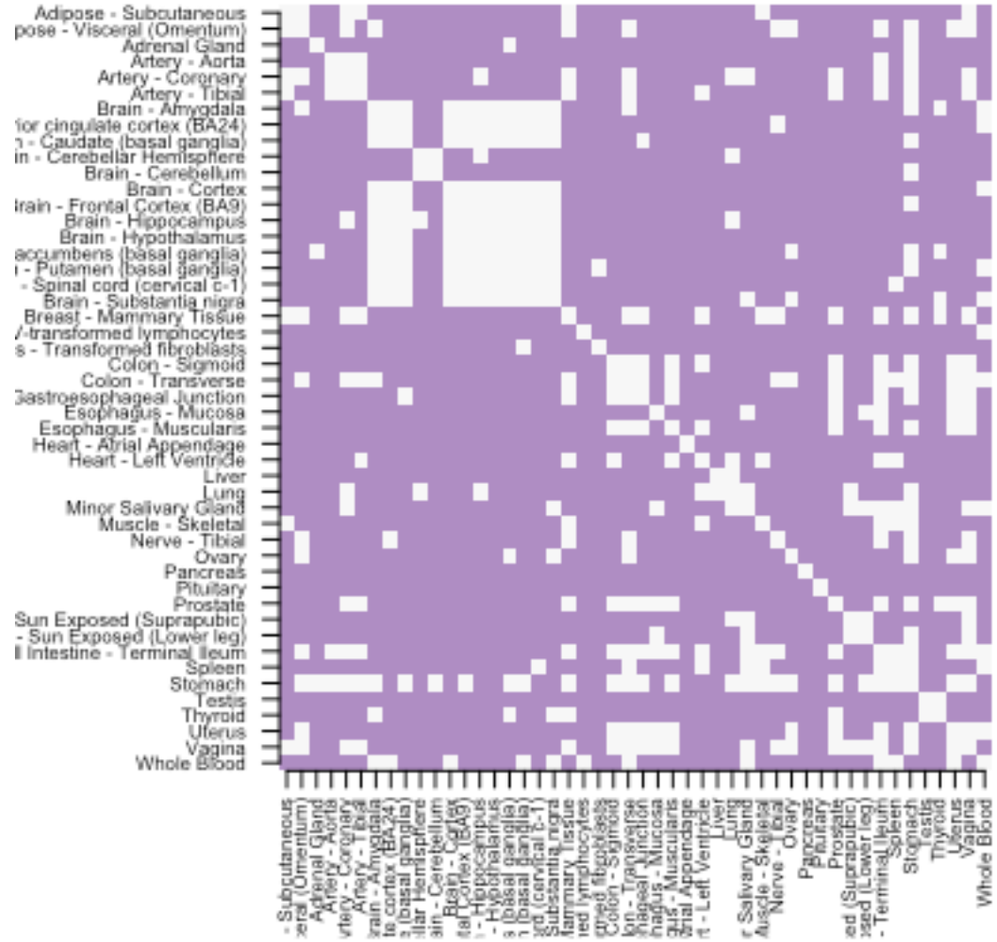
Which method separates tissues better?

separates ■ ☐ does not separate

Hierarchical



Grade of Membership (GoM)

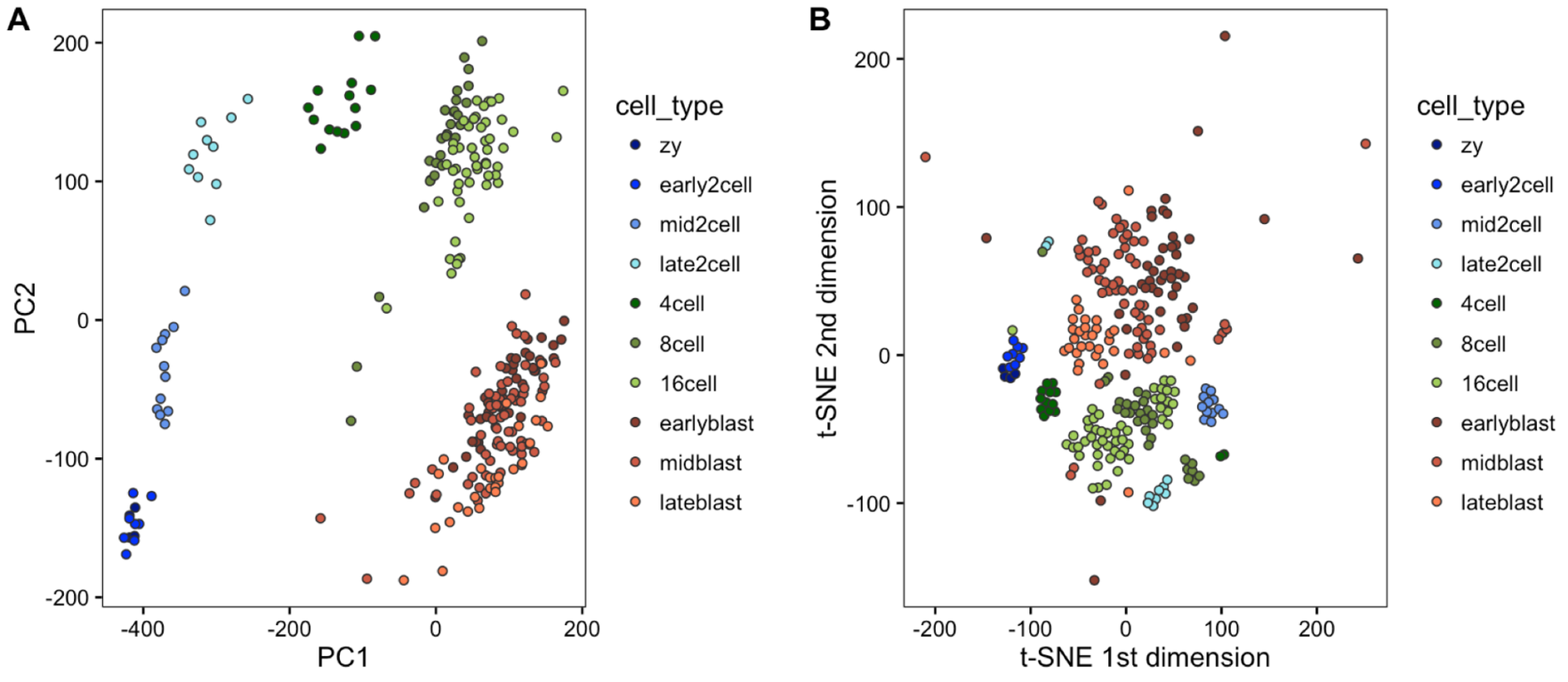


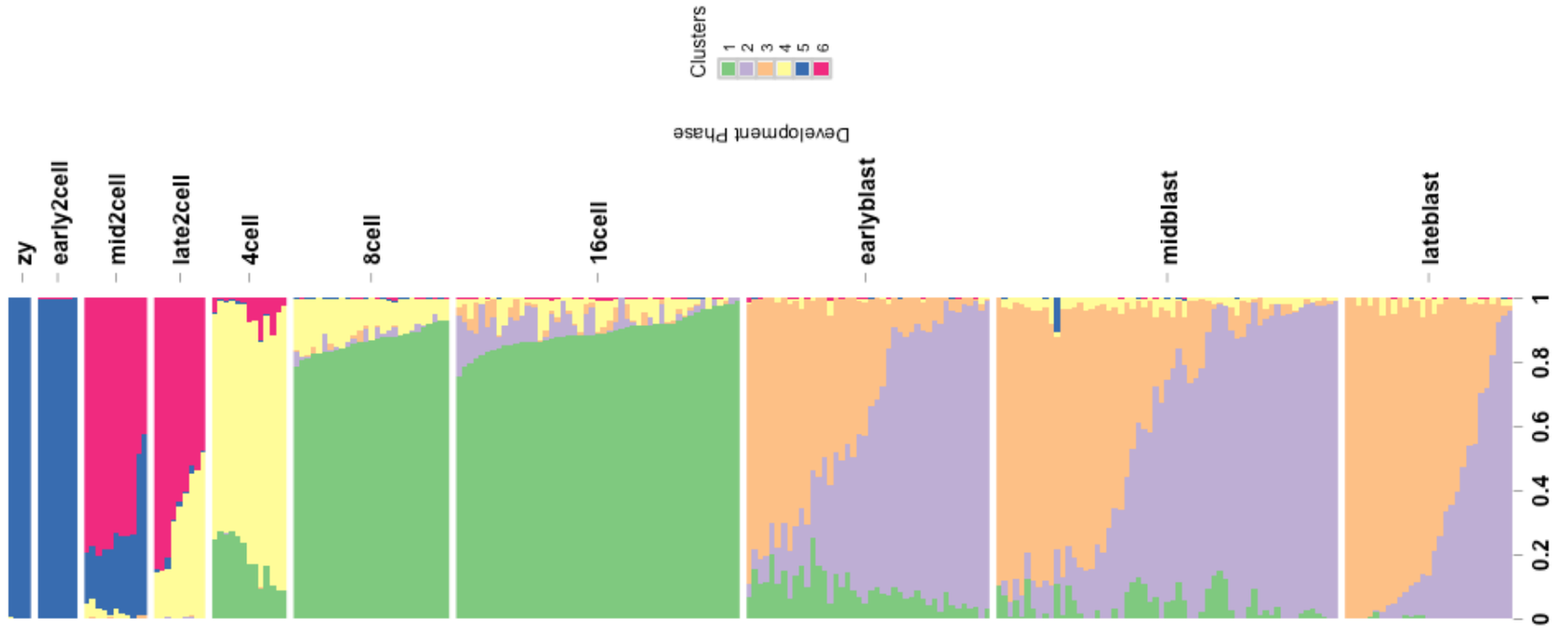
Single cell Example

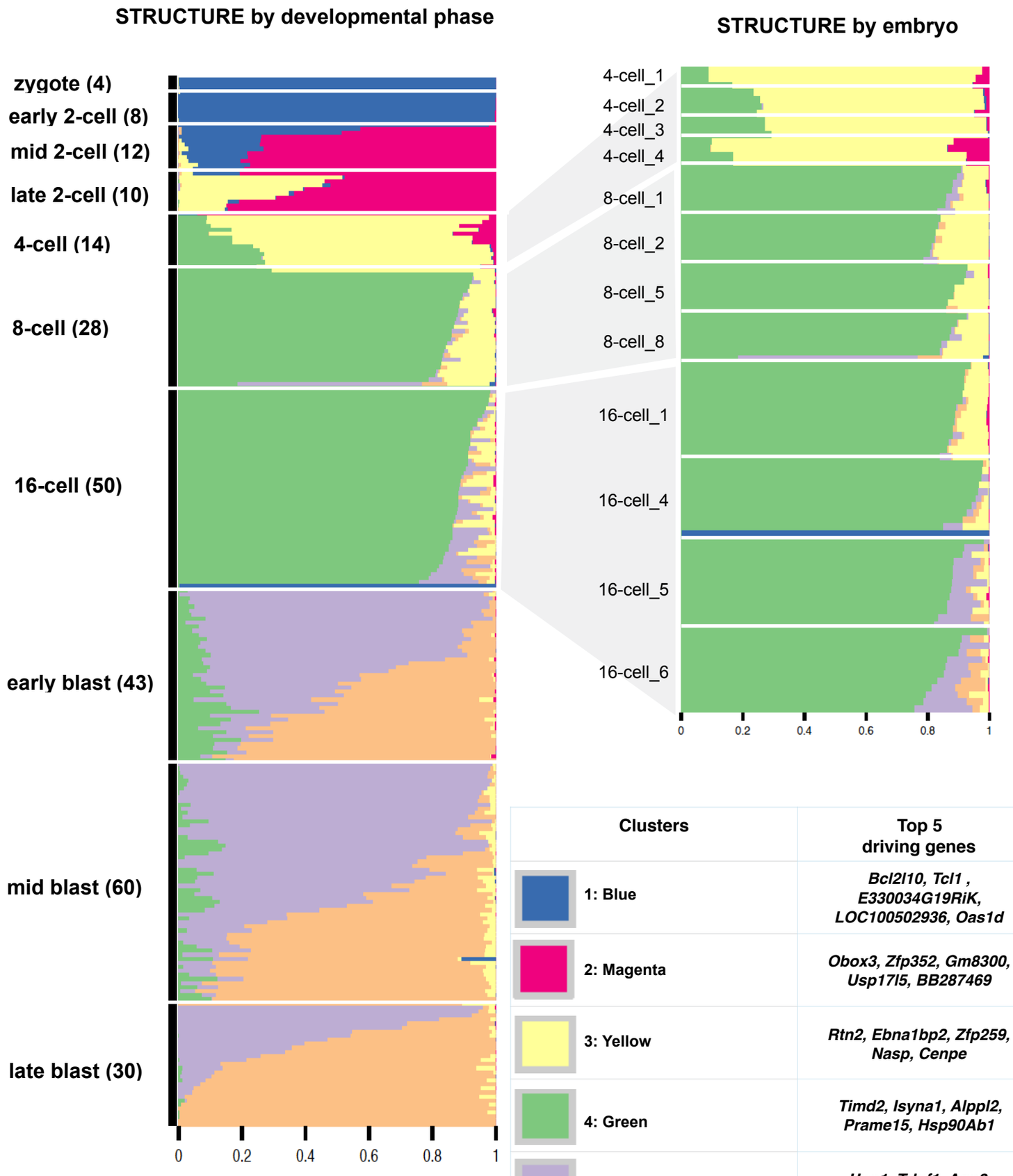
Deng et al (2014) collected single cell data
across different developmental stages from zygote,
2-cell, 4-cell, 8-cell, 16-cell, and three stages
of blastocysts (early, mid and late)

We plot admixture plot along with other dimension
reduction methods like PCA and t-SNE

Deng Q1, Ramsköld D, Reinius B, Sandberg R. *Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells.*
Science. 2014 Jan 10;343(6167):193-6


A
B
PC2
PC1
t-SNE 2nd dimension
t-SNE 1st dimension
cell_type
zy
early2cell
mid2cell
late2cell
4cell
8cell
16cell
earlyblast
midblast
lateblast


Clusters
1
2
3
4
5
6
Development Phase
zy
early2cell
mid2cell
late2cell
4cell
8cell
16cell
earlyblast
midblast
lateblast
0
0.2
0.4
0.6
0.8
1



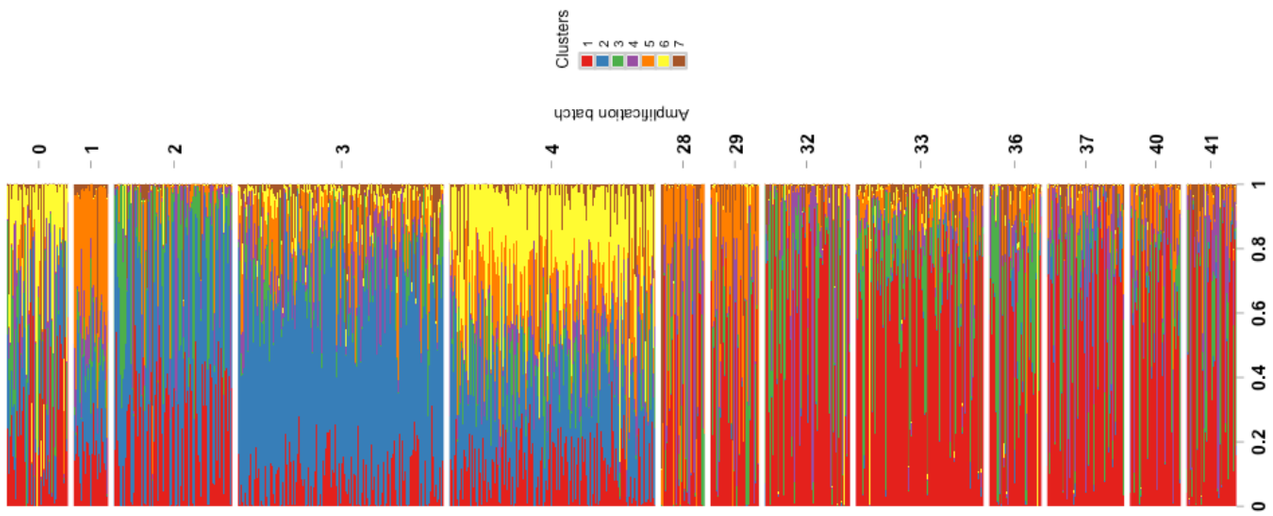
Clusters	Top 5 driving genes	GO annotations summary
1: Blue	*Bcl2l10, Tcl1 , E330034G19RiK, LOC100502936, Oas1d*	*Gamete generation, Oocyte development & differentiation, Cell devision regulation*
2: Magenta	*Obox3, Zfp352, Gm8300, Usp17l5, BB287469*	*Nuclear body, Centriole, Microtubule organizing center 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, Metabolic Process, Blood Vessel Morphogenesis*
6: Orange	*Actb, Krt18, Fabp3, Id2, Tspan8*	*Extracellular Exome/Organelle, Membrane Bound Organelle, Actin Cytoskeleton, Actin Filament Bundle*

Discussion

The results from the clustering method may be driven by partly by batch effects, besides the biological effects.

For data where the batches are confounded with biological subgroups, it becomes difficult to separate the two.

We encountered this issue in the analysis of *Jaitin et al (2014)* scRNA-seq data from mouse spleen.



DA Jaitin, E Kenigsberg et al. *Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types.* Science 14 Feb 2014:Vol. 343, Issue 6172, pp. 776-779

Discussion

The method presented is unsupervised.

However, many recent studies sequence single cells from different cell types. These single cell information can be used to drive the clustering for bulk-RNA data.

In such a case, the clusters obtained would conform more with the cell-types by design and that was our primary interest.

Manuscript: **Clustering RNA-seq expression data using grade of membership models (**http://biorxiv.org/content/early/2016/05/03/051631**)**

R Package: *CountClust*

Bioconductor release version
(**https://www.bioconductor.org/packages/3.3/bioc/html/CountClust.html**)

Under development version
(**https://github.com/kkdey/CountClust**)

Acknowledgements

Matt Taddy (UChicago Booth School),
Jonathan Pritchard (Stanford University),
Yoav Gilad (UChicago Human Genetics)
Amos Tanay, Effi Kenigsberg,
Po Yuan Tung, John Blischak (Gilad Lab)
Gao Wang, Raman Shah, Nan Xiao (Stephens Lab)

Co-authors

Joyce Hsiao

Matthew Stephens