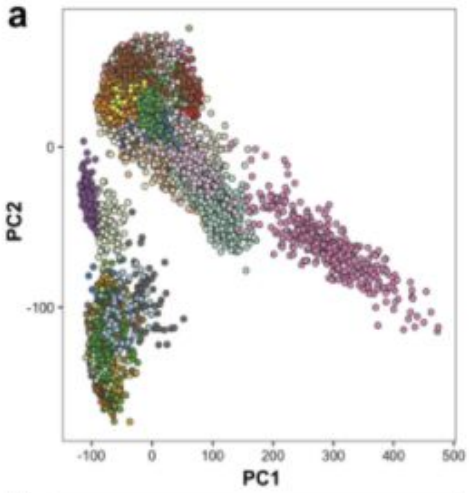
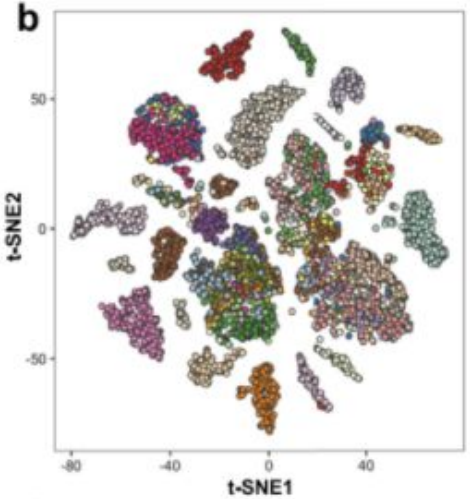
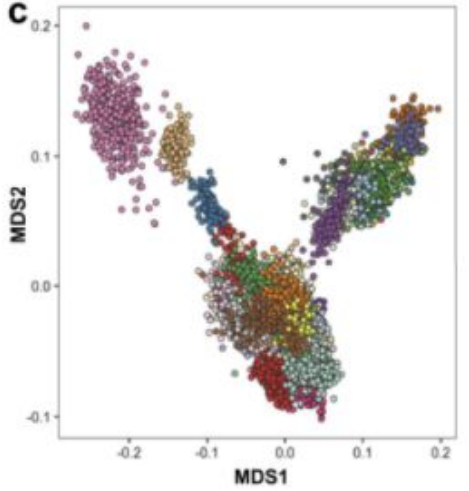
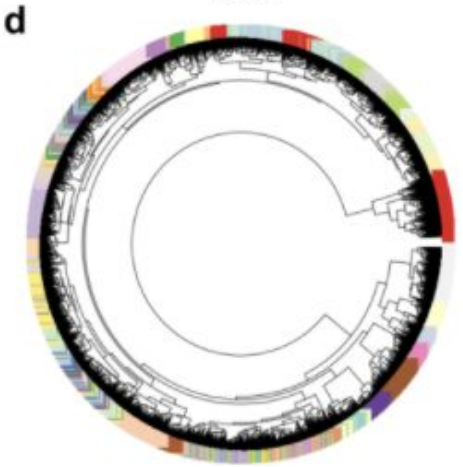








Visualizing the structure of RNA-seq expression data using grade of membership models

Kushal K. Dey, Chiaowen Joyce Hsiao, Matthew Stephens

Background

- ➔ “Grade of Membership models” = “Admixture models” = “topic models” = “Latent Dirichlet Allocation”
  - ◆ Generalization of cluster models that allow each word in a document to be a member of multiple clusters
  - ◆ Generates clusters based on word counts that appear in documents
- ➔ For our purposes:
  - ◆ Cluster → Topic (ex. Immune System)
  - ◆ Document → Sample (ex. Lung tissue)
  - ◆ Word → RNA sequence from a gene (ex. *myh6*) ✧

Introduction

- ➔ Clustering of genes and samples have been widely in the interpretation of genetic data
  - ◆ Clustering of genes can identify genes that are working together/coregulated
  - ◆ Clustering of samples can identify biologically distinct subgroups of a population
- ➔ This paper focuses on cluster analysis of samples rather than the clustering of genes
- ➔ Traditional clustering methods attempt to partition samples into distinct groups that show similar expression patterns
  - ◆ Problematic; typical gene expression data can be too complex to be captured by such partitioning ✧

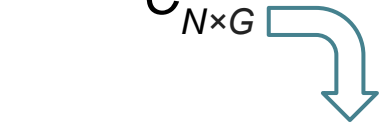
Introduction

- ➔ Using an LDA model allows for each sample to have partial membership to multiple topics
  - ◆ Each sample has a “grade” of membership in each cluster
    - Where “GoM” model name comes from
- ➔ Generally, GoM is used for “cell type deconvolution”
  - ◆ Uses cell-specific expression profiles to estimate proportions of cell types in a mixture
  - ◆ Can be generalized to interpretation of RNA-sequence data ✧

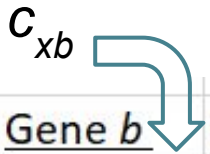
Notation

- RNA-seq data of *N* samples has been summarized by a table of counts
  - $C_{NxG} = c_{ng}$, where c_{ng} is the number of reads from sample *n* mapped to gene *g*
- It is assumed that each sample *n* has some proportion of its reads, q_{nk} , coming from topic *k*
  - Some of the genes expressed from each sample corresponds to a specific topic
- Each topic *k* is characterized by a probability vector, θ_k , whose g^{th} element represents the relative expression of gene *g* in topic *k* ✧

$C_{N \times G}$



		Gene *a*	Gene *b*	Gene *c*	Topic 1	Topic 2	Topic 3
	Sample *x*	5	6	3	0.4	0.4	0.2
	Sample *y*	2	1	9	0.2	0.1	0.7
	Sample *z*	5	6	4	0.35	0.35	0.3
(Metabolism)	Topic 1	0.50	0.40	0.10			
(Digestion)	Topic 2	0.40	0.50	0.10			
(Immunity)	Topic 3	0.15	0.15	0.70			



		Gene *a*	Gene *b*	Gene *c*	Topic 1	Topic 2	Topic 3
	Sample *x*	5	6	3	0.4	0.4	0.2
	Sample *y*	2	1	9	0.2	0.1	0.7
	Sample *z*	5	6	4	0.35	0.35	0.3
(Metabolism)	Topic 1	0.50	0.40	0.10			
(Digestion)	Topic 2	0.40	0.50	0.10			
(Immunity)	Topic 3	0.15	0.15	0.70			

		Gene *a*	Gene *b*	Gene *c*	Topic 1	Topic 2	Topic 3
	Sample *x*	5	6	3	0.4	0.4	0.2
	Sample *y*	2	1	9	0.2	0.1	0.7
	Sample *z*	5	6	4	0.35	0.35	0.3
(Metabolism)	Topic 1	0.50	0.40	0.10			
(Digestion)	Topic 2	0.40	0.50	0.10			
(Immunity)	Topic 3	0.15	0.15	0.70			



		Gene *a*	Gene *b*	Gene *c*	Topic 1	Topic 2	Topic 3
	Sample *x*	5	6	3	0.4	0.4	0.2
	Sample *y*	2	1	9	0.2	0.1	0.7
	Sample *z*	5	6	4	0.35	0.35	0.3
(Metabolism)	Topic 1	0.50	0.40	0.10			
(Digestion)	Topic 2	0.40	0.50	0.10			
(Immunity)	Topic 3	0.15	0.15	0.70			



input

		Gene *a*	Gene *b*	Gene *c*	Topic 1	Topic 2	Topic 3
	Sample *x*	5	6	3	0.4	0.4	0.2
	Sample *y*	2	1	9	0.2	0.1	0.7
	Sample *z*	5	6	4	0.35	0.35	0.3
(Metabolism)	Topic 1	0.50	0.40	0.10			
(Digestion)	Topic 2	0.40	0.50	0.10			
(Immunity)	Topic 3	0.15	0.15	0.70			

		Gene *a*	Gene *b*	Gene *c*
	Sample *x*	5	6	3
	Sample *y*	2	1	9
	Sample *z*	5	6	4
(Metabolism)	Topic 1	0.50	0.40	0.10
(Digestion)	Topic 2	0.40	0.50	0.10
(Immunity)	Topic 3	0.15	0.15	0.70

Topic 1	Topic 2	Topic 3
0.4	0.4	0.2
0.2	0.1	0.7
0.35	0.35	0.3

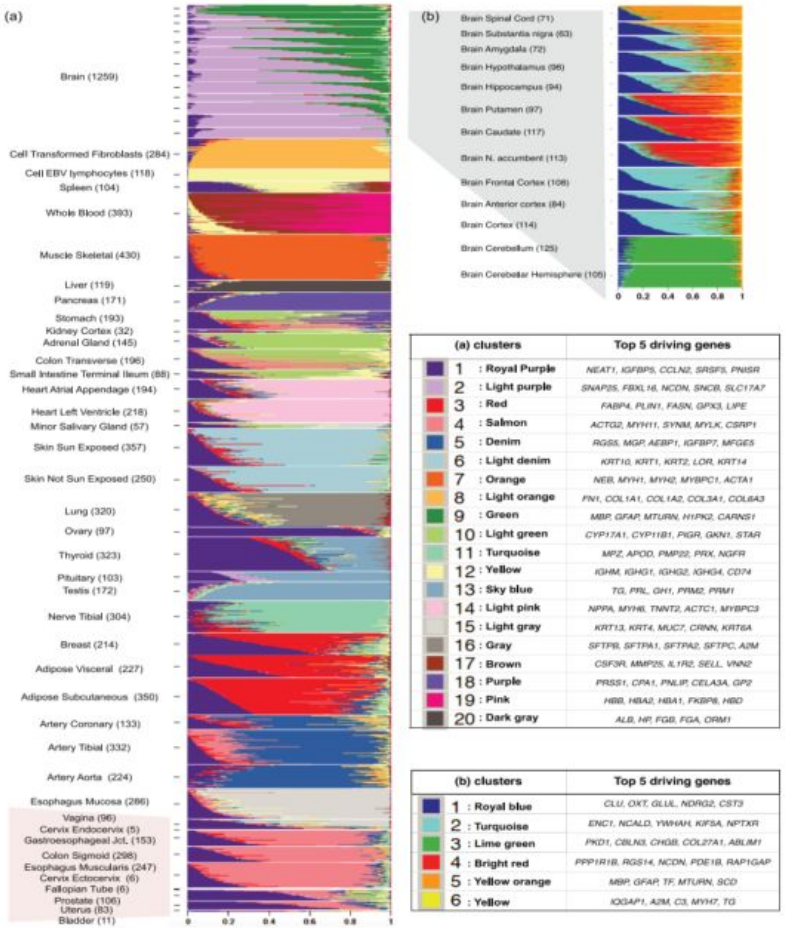
outputs

		Gene *a*	Gene *b*	Gene *c*		Topic 1	Topic 2	Topic 3
	Sample *x*	5	6	3		0.4	0.4	0.2
	Sample *y*	2	1	9		0.2	0.1	0.7
	Sample *z*	5	6	4		0.35	0.35	0.3
(Metabolism)	Topic 1	0.50	0.40	0.10				
(Digestion)	Topic 2	0.40	0.50	0.10				
(Immunity)	Topic 3	0.15	0.15	0.70				

Discerned after

Methods

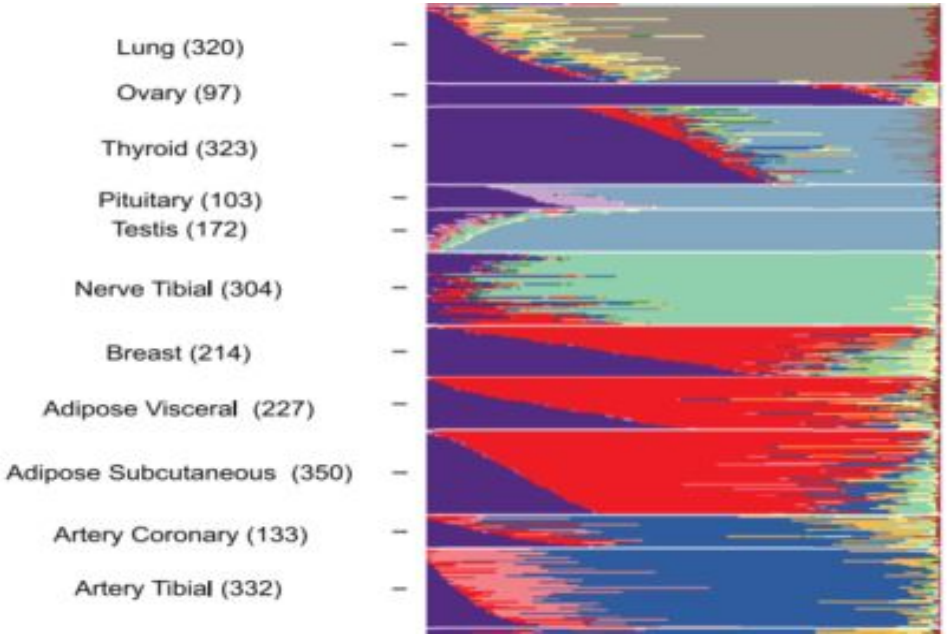
- ➔ The number of topics *k* was set by the authors (*k* = 20)
- ➔ Data used: 16,069 genes from 8,555 tissue samples from 53 tissues in GTEx project database
- ➔ Fitting the the GoM model to the data resulted in the estimated …
  - ◆ topic membership proportions *q* for each sample
  - ◆ expression values *θ* for each topic
- ➔ Membership proportions were visualized using a “structure plot”
  - ◆ Represents the estimated membership proportions of each sample as a stacked bar chart, with bars of different colors representing different topics
  - ◆ Samples that have similar membership proportions have similar amounts of each colour



(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, HIPK2, 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, KIFSA, 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, MYH7, TG*

Fig 1. GTEx tissues Structure plot. (a): Structure plot of estimated membership proportions for GoM model with K = 20 clusters fit to 8, 555 tissue samples from 53 tissues in GTEx data. Each horizontal bar shows the cluster membership proportions for a single sample, ordered so that samples from the same tissue are adjacent to one another. Within each tissue, the samples are sorted by the proportional representation of the underlying

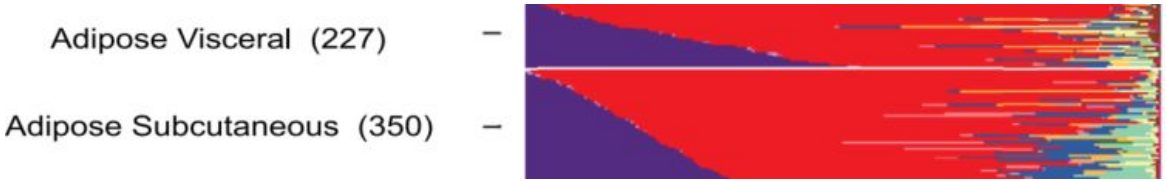


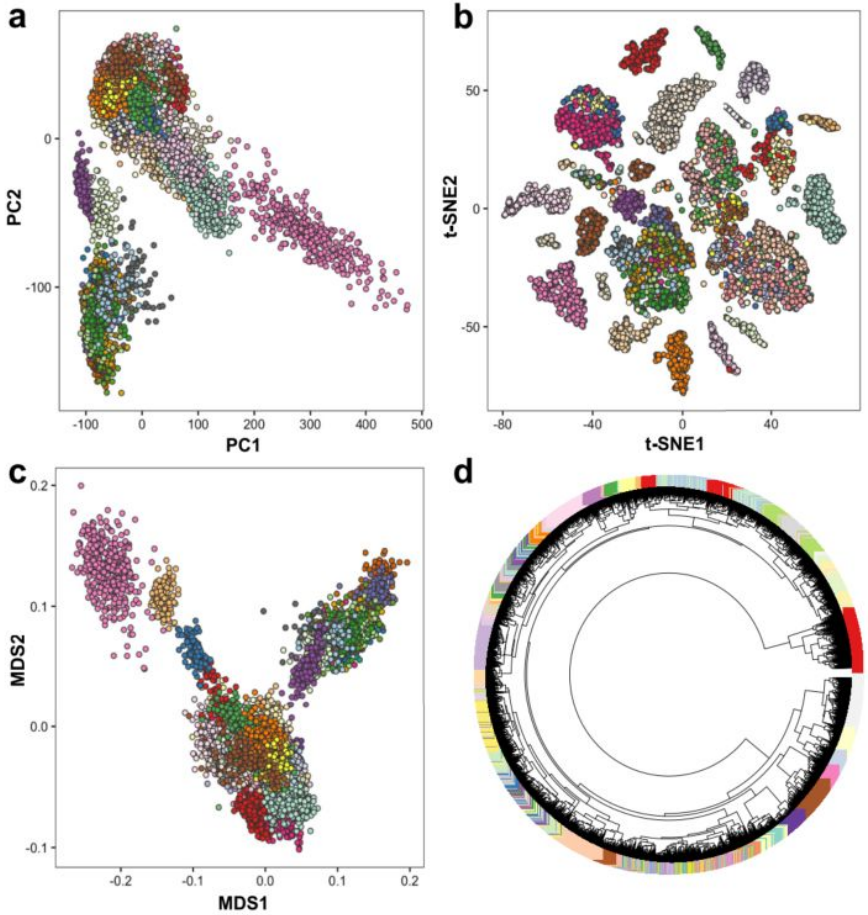
(a) clusters	Top 5 driving genes
1 : Royal Purple	*NEAT1, IGFBP5, CCLN2, SRSF5, PNISR*
2 : Light purple	*SNAP25, FBXL16, NCDN, SNCB, SLC17A7*
3 : Red	*FABP4, PLIN1, FASN, GPX3, LIPE*
4 : Salmon	*ACTG2, MYH11, SYNM, MYLK, CSRP1*
5 : Denim	*RGS5, MGP, AEBP1, IGFBP7, MFGE5*
6 : Light denim	*KRT10, KRT1, 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*

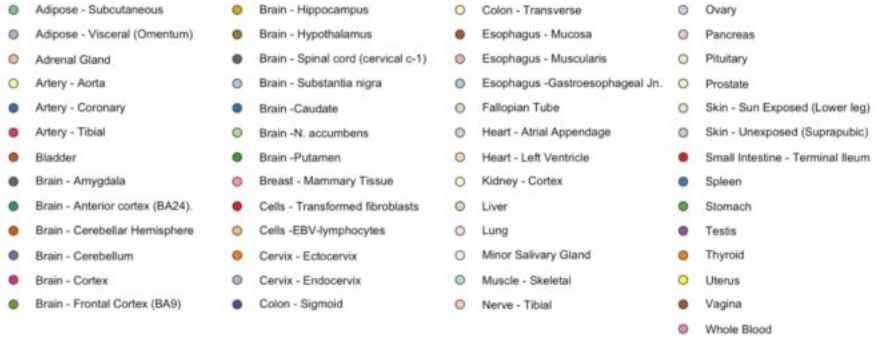
- ➔ Structure plot of estimated membership proportions with k = 20 topics to fit 8,555 samples from 53 tissues
- ➔ Each horizontal bar shows topic membership proportions for single sample

Results

- ➔ Increasing *k* seems to highlight finer structure in the data
  - ◆ Authors tested results where *k* = 5, 10, 15, 20
- ➔ Some tissues are represented by essentially a single cluster/colour (ex. Pancreas, liver), while others are represented by a mixture (ex. Thyroid, Spleen)
- ➔ GoM results highlight biological similarity by assigning similar membership proportions to samples from those tissues:







A. PCA plot of PC1 vs PC2
B. Plotted first two dimensions of *t*-SNE plot
C. Plotted first two dimensions of the MDS plot
D. Dendrogram of the hierarchical clustering of the tissue samples based on the log CPM expression data (with average linkage and euclidean distance)

➔ GoM model is better at visualizing data than PCA and MDS

Results

- ➔ Samples clustered by tissue rather than other factors, speaking well for data quality
  - ◆ Could have also clustered by technical variables, such as sequencing batch or sample collection center
- ➔ GoM provides estimate of the mean expression of each gene making it straightforward to see which genes are responsible for cluster differences
- ➔ GoM allows for partial membership in each cluster, so it is better able to highlight partial similarities among distinct tissues
  - ◆ Sky blue cluster appears in testis, pituitary, and thyroid, reflecting shared hormonal-related processes ✧

Auxiliary Experiments

- ➔ Sub-analysis of brain tissues
  - ◆ GoM applied to brain tissue samples, k = 6, found further intricacies
- ➔ Single-cell RNA-seq data
  - ◆ Applied GoM to:
    - Mouse spleen data
      - Most cells were assigned membership to several clusters, not ideal
      - Data seemed to be strongly correlated with sequencing batch
    - Mouse preimplantation embryo data
      - Highlights changes in expression profiles that occur through early embryonic development stages
      - Enrichment analysis shows these changes reflect biological processes ✧

Discussion

- ➔ Goal of the paper was to show that GoM models are valuable for visualizing RNA-seq data structure
- ➔ Authors provide tools that identify which genes are most distinctively expressed in each topic
- ➔ Tool has potential to:
  - ◆ Highlight issues for investigation
  - ◆ Identify abnormalities
  - ◆ Generate hypotheses for future testing ✧

Discussion

- ➔ Different methods have different strengths and weaknesses, meaning they can complement each other
  - ◆ *t*-SNE provides clearer indication of the cluster structure than PCA, PCA does better for capturing ordering of developmental samples for mice embryos
- ➔ Fitting GoM models can be computationally intensive for larger datasets
  - ◆ Using the GTEx data (n = 8,555, K = 20) took 56 hours!
- ➔ There is unlikely to be a “true” value of K in practice, so it may be useful to evaluate many and watching the fitted model evolve
- ➔ This being said, structure plots with $K > 30$ get hard to interpret as there are only so many distinguishable colours! ✧

R Packages Used

`CountClust` → Author's Bioconductor package for implementing methods

`maptpx` → Fitting the GoM model