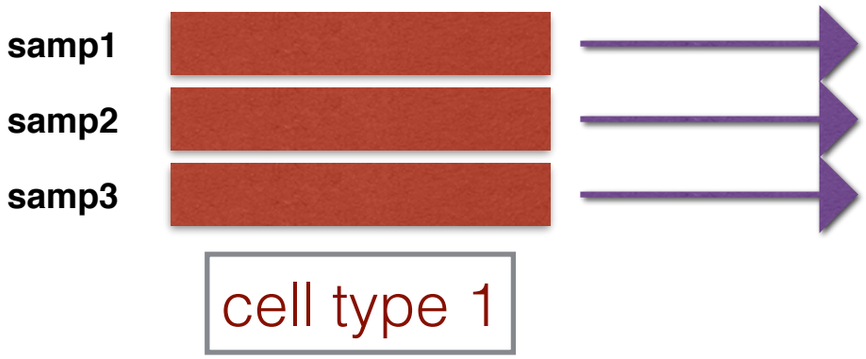
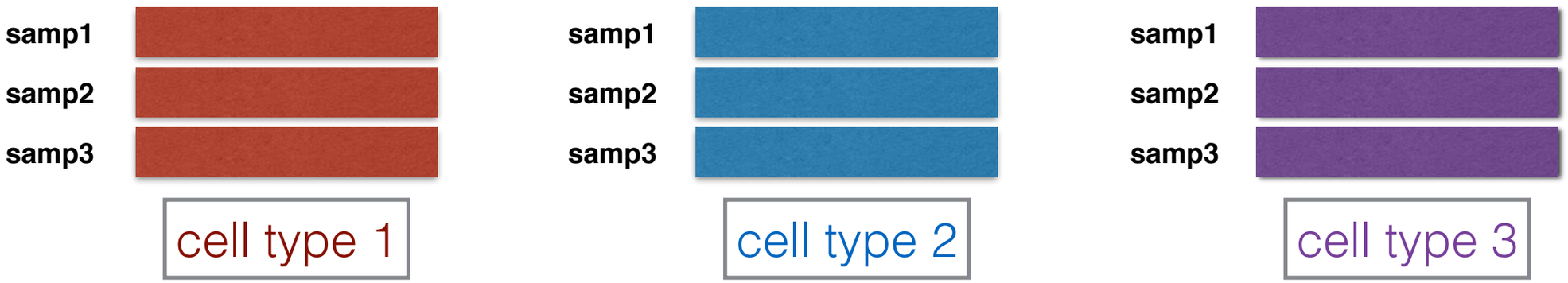
Grade of membership classification in RNA-seq data

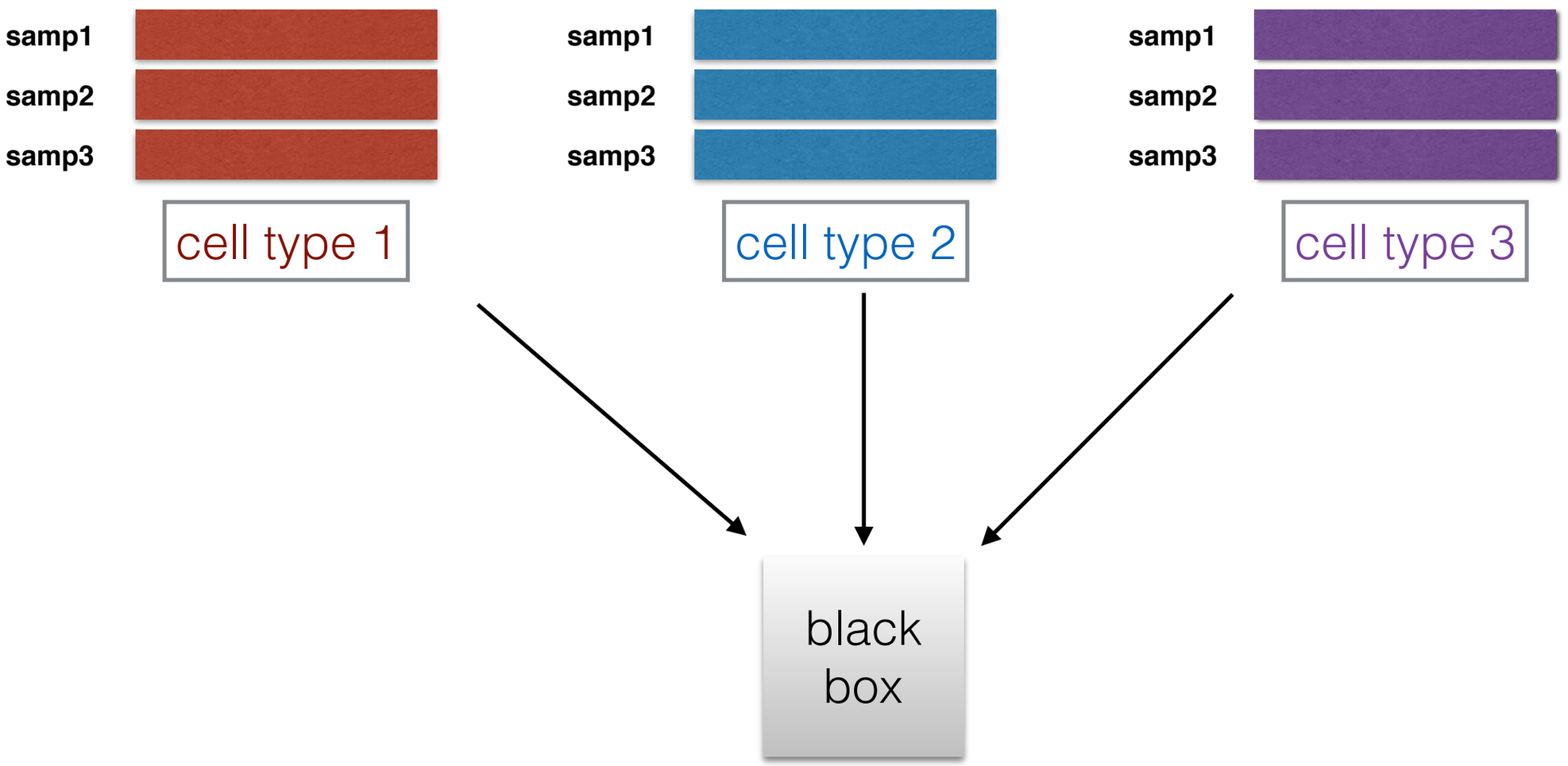
- Model intuition
- Statistical framework
- Workflow for RNA-seq data
- GTEx applications
- Single-cell applications
- Summary

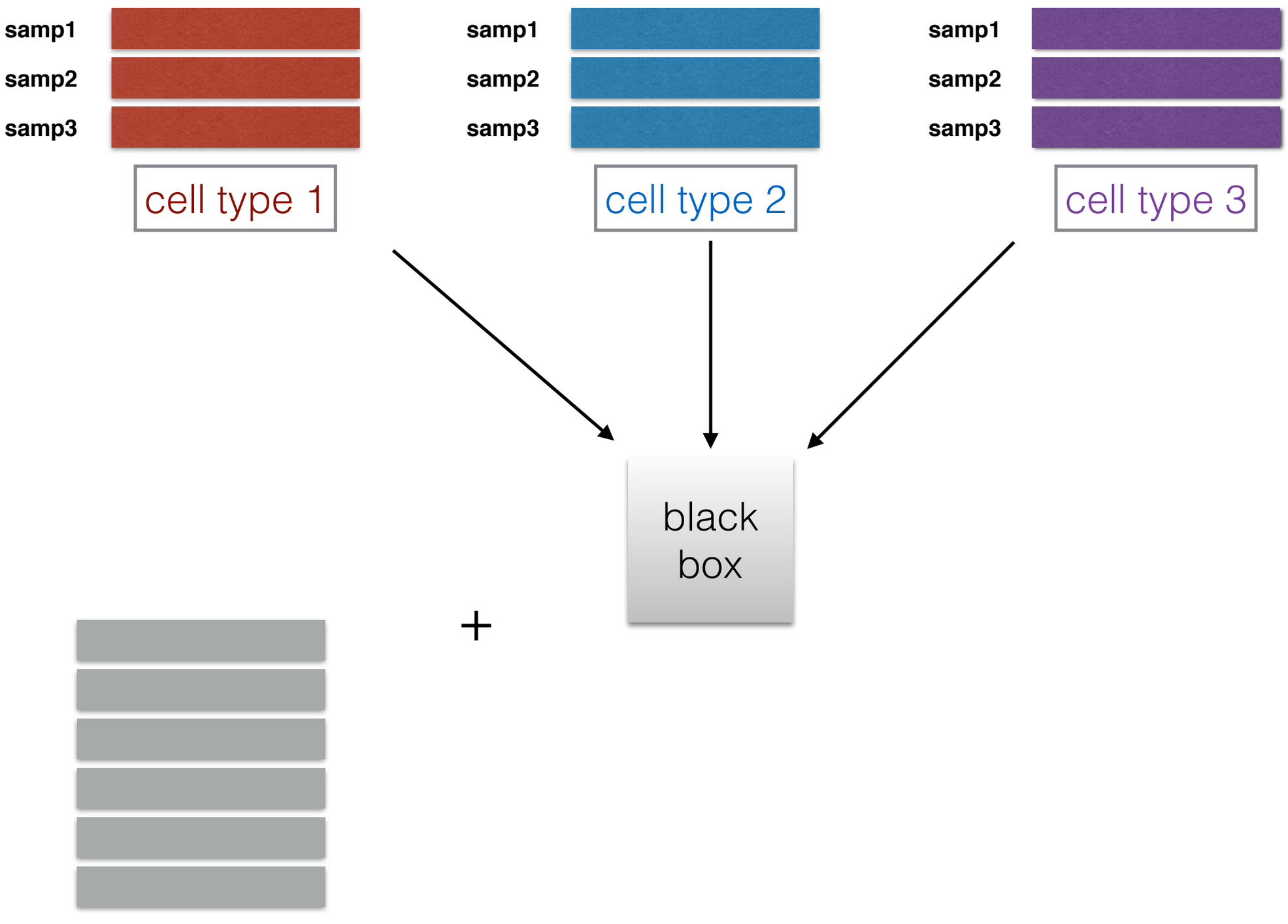


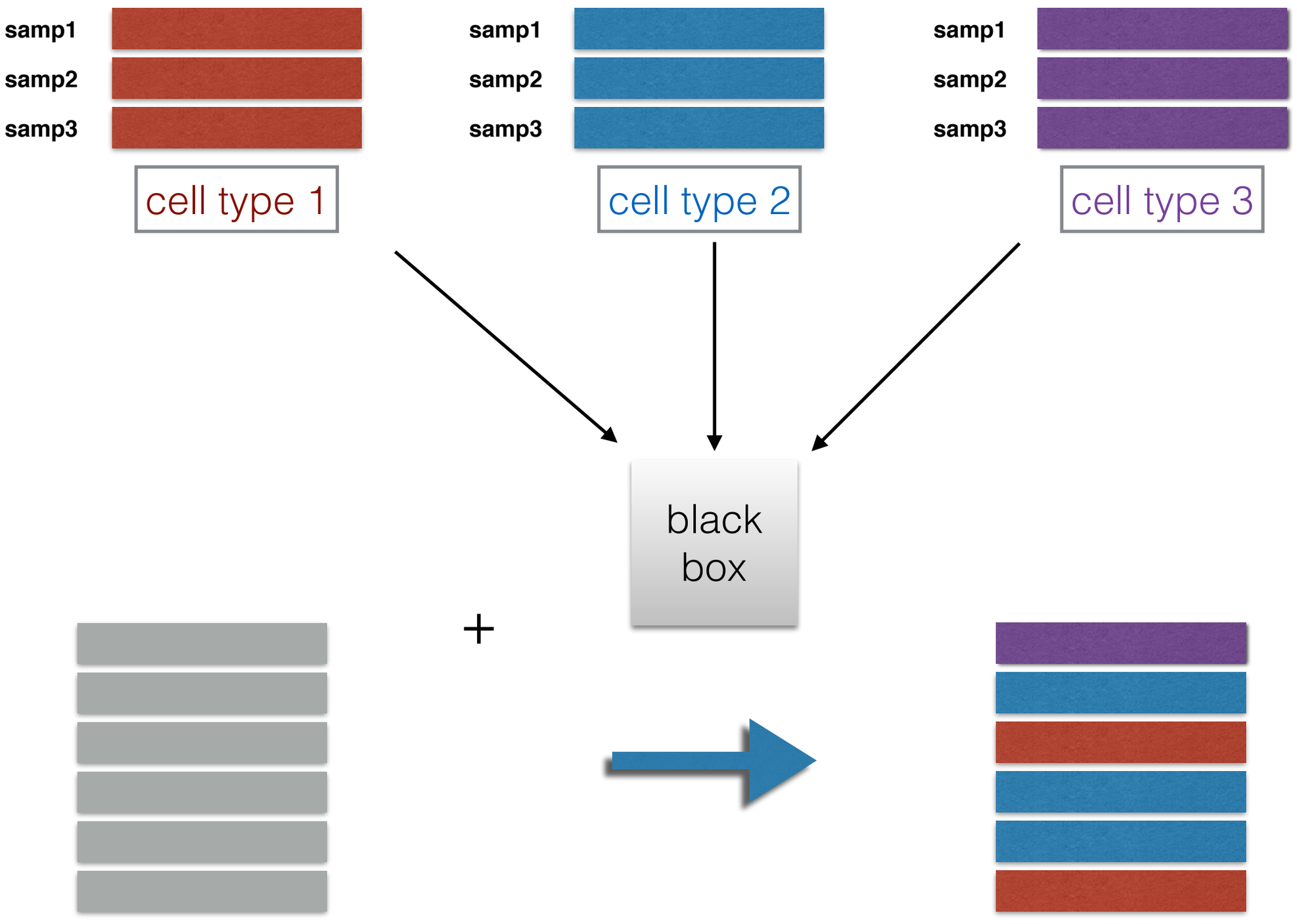
$$(c_{11}, c_{12}, c_{13}, \cdots, c_{1G})$$
$$(c_{21}, c_{22}, c_{23}, \cdots, c_{2G})$$
$$(c_{31}, c_{32}, c_{33}, \cdots, c_{3G})$$

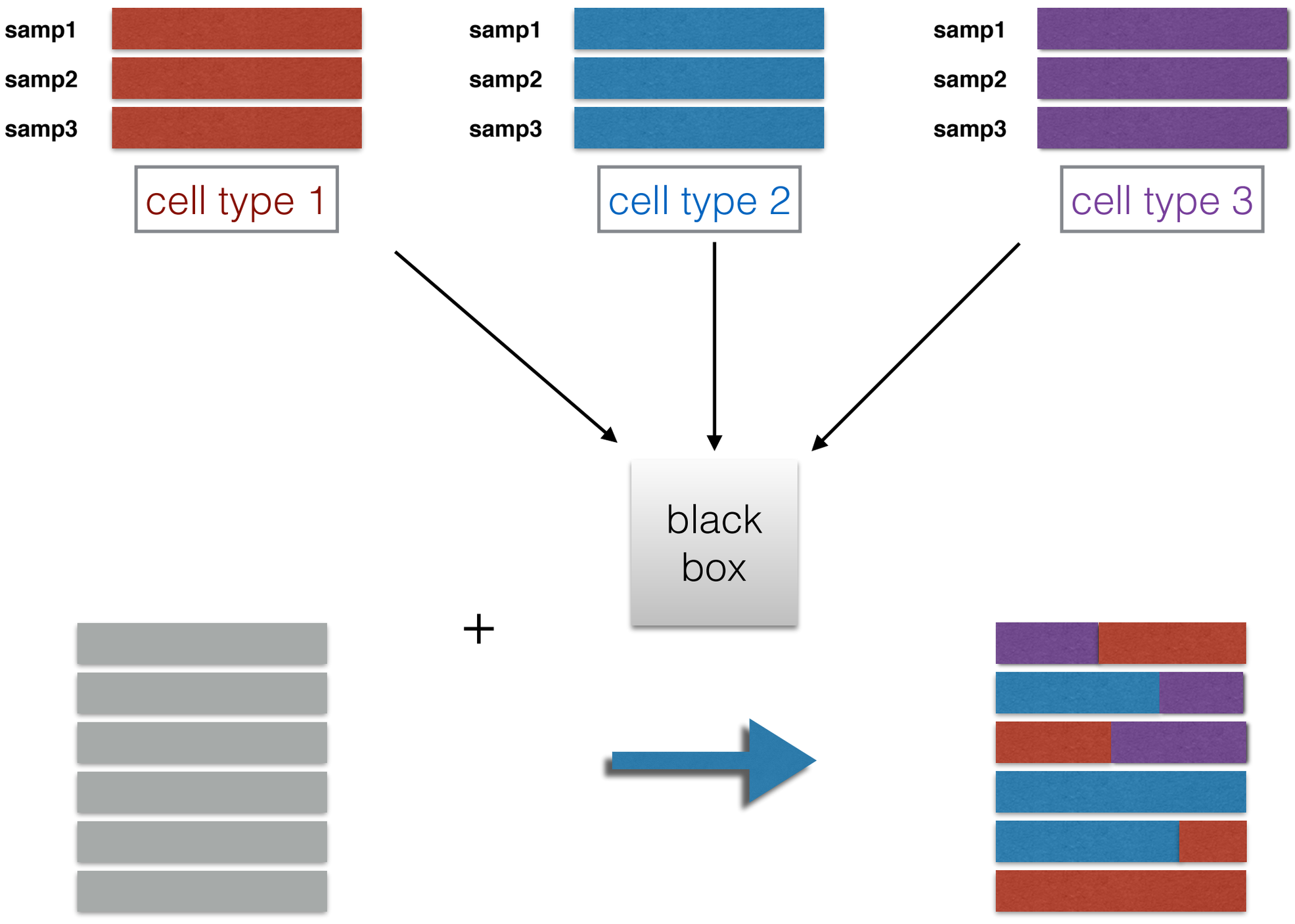
RNA-seq reads count data across G genes for each sample


samp1
samp2
samp3
cell type 1
samp1
samp2
samp3
cell type 2
samp1
samp2
samp3
cell type 3


samp1
samp2
samp3
cell type 1
samp1
samp2
samp3
cell type 2
samp1
samp2
samp3
cell type 3
black box


samp1
samp2
samp3
cell type 1
samp1
samp2
samp3
cell type 2
samp1
samp2
samp3
cell type 3
black box
+


samp1
samp2
samp3
cell type 1
samp1
samp2
samp3
cell type 2
samp1
samp2
samp3
cell type 3
black box
+

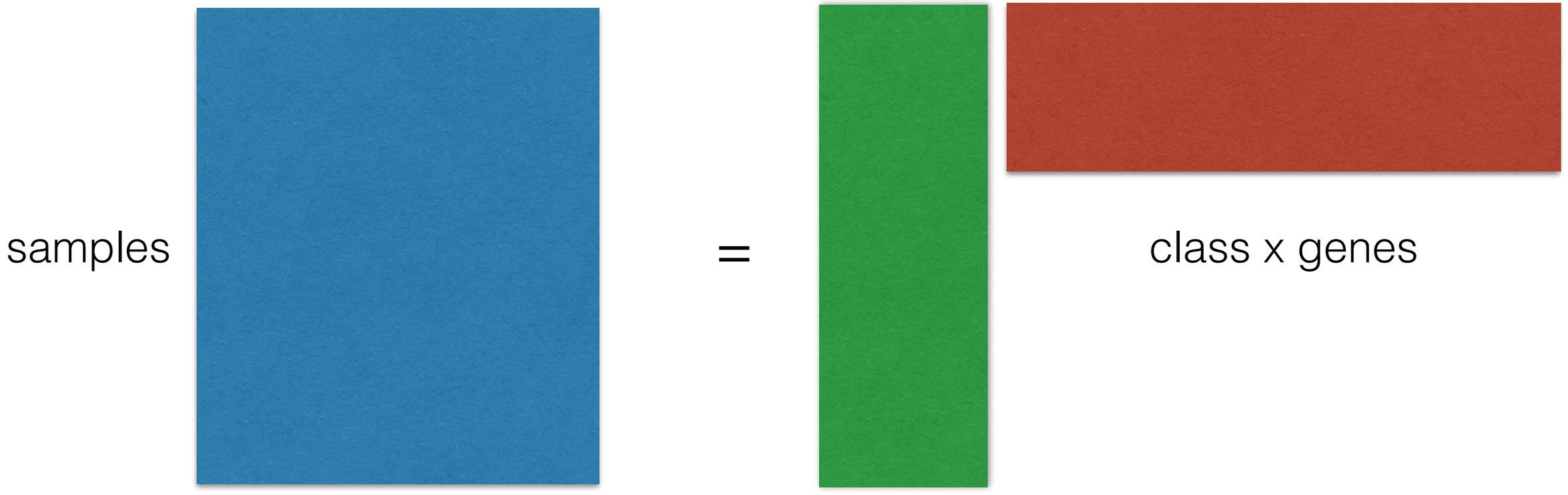

samp1
samp2
samp3
cell type 1
samp1
samp2
samp3
cell type 2
samp1
samp2
samp3
cell type 3
black box
+

Grade of membership model

The model

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

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



Unsupervised approach

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

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

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

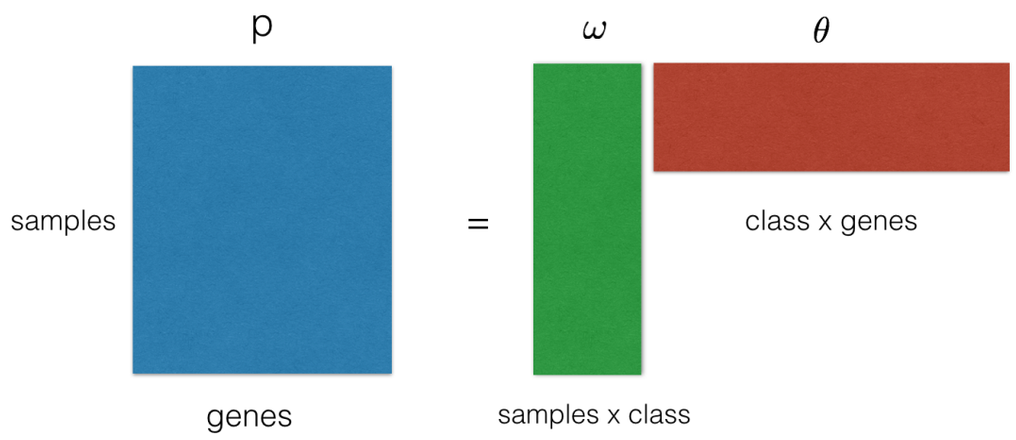
ω_{nk} and θ_{kg} were determined in an unsupervised manner assuming Dirichlet priors

Note: R package **maptpx** was used to perform unsupervised GoM fitting.

classtpx: an semi-supervised approach

In ***omega.fix***, some samples coming from each cell type *k* are treated as training and we assign grade of membership of those samples in class *k* to be 1 and that in other classes to be 0.

In ***theta.fix*** method, we fix $\theta_{k.}$ for cluster *k* by pooling the relative expression patterns across genes for training samples coming from cluster *k*



The test data would be actual tissue samples which would be a mixture of these different cell types or classes.

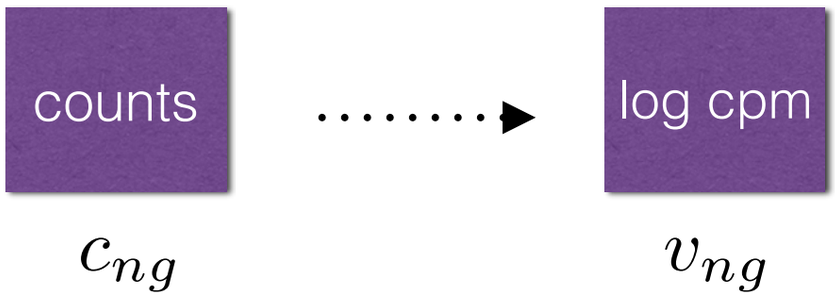
RNA-seq application

Challenges

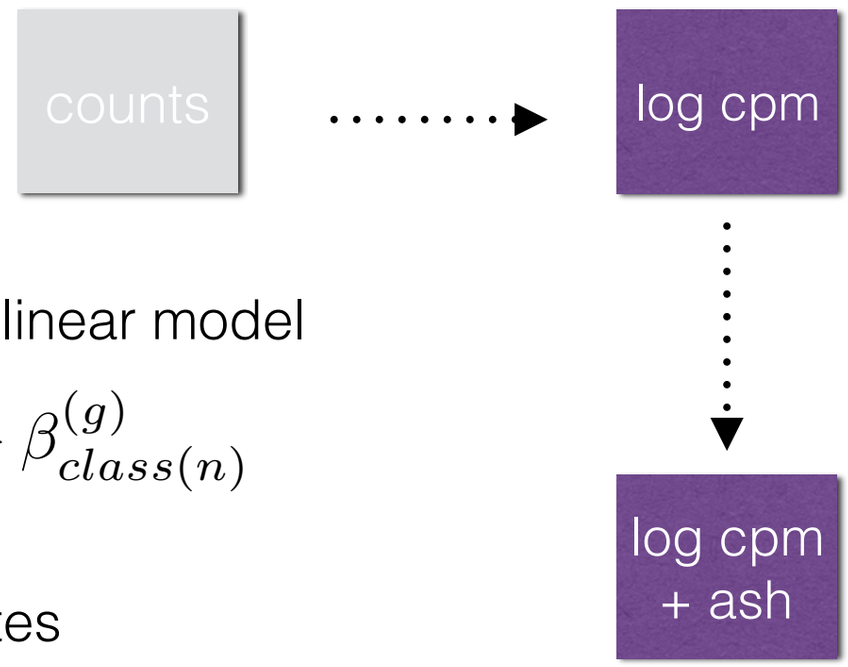
- Normalization
- Expression noise (should we shrink?? adaptively ?)
- Count-based or log-CPM based analysis?

Workflow

$$c_{ng} \rightarrow log_2 \left(\frac{c_{ng} + 0.5}{c_{n+} + 1} \times 10^6 \right)$$



Workflow



1. Fit a Bayesian linear model

$$v_{ng} \sim \mu^{(g)} + \beta^{(g)}_{class(n)}$$

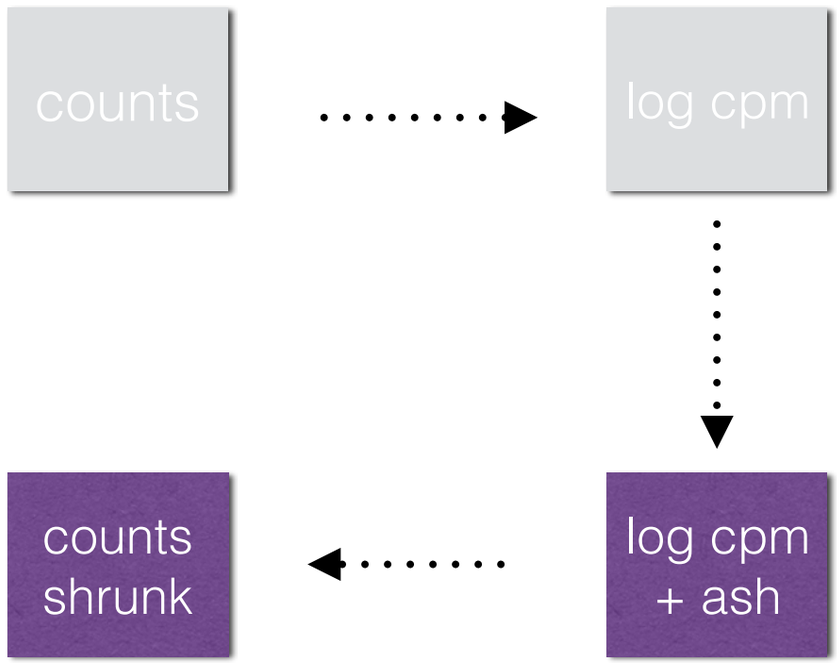
2. Obtain estimates

$$\hat{\beta}^{(g)}_{class(n)}, se(\hat{\beta}^{(g)}_{class(n)})$$

3. Shrink $\hat{\beta}^{(g)}_{class(n)}$ to 0 (ash, *Stephens 2016)*

$$v^{ash}_{ng} := \hat{\mu}^{(g)} + \hat{\beta}^{(g),ash}_{class(n)}$$

Workflow



4. Reverse transform

$$v_{ng}^{ash} \overset{rev.voom}{\rightarrow} c_{ng}^{ash}$$

Classification in RNA-seq data

1. Model-based versus non-model based
2. Soft versus hard classification

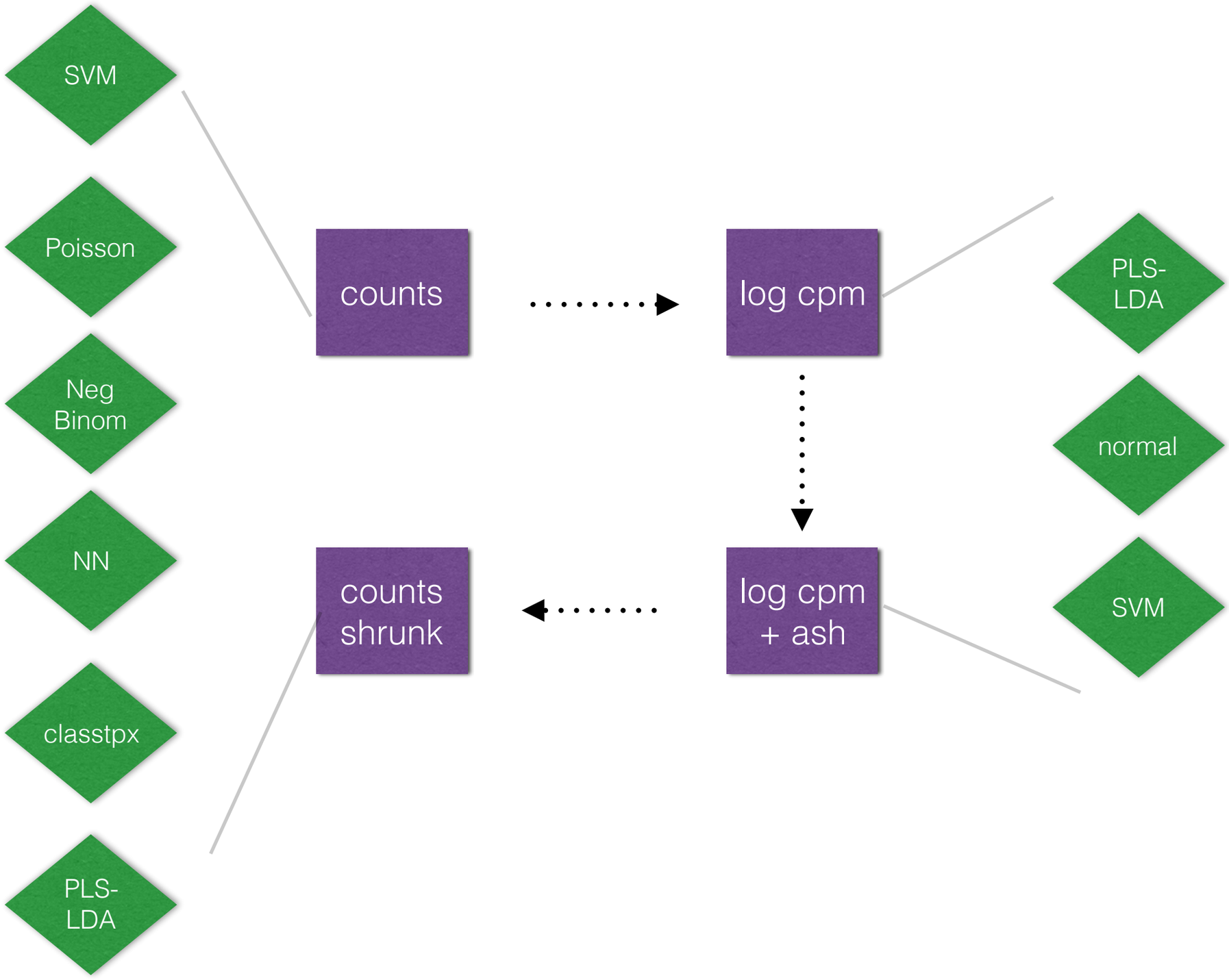
- Non-model based approach:

  Support Vector machines, PLS-LDA, Neural network

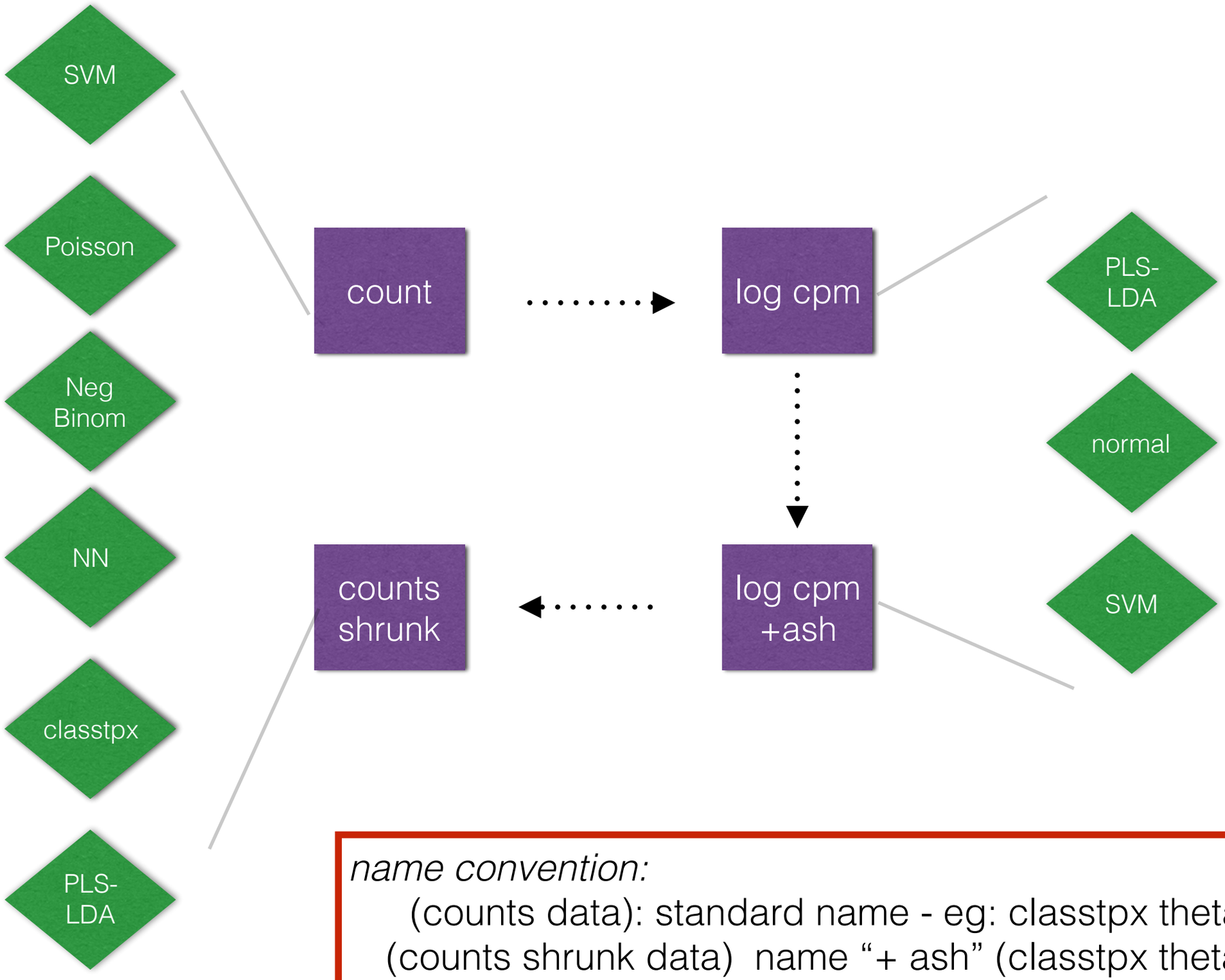
- Model based approach

  Poisson model on counts, Negative Binomial model on counts, Normal model, ***grade of membership approach***

Workflow


SVM
Poisson
Neg Binom
NN
classtpx
PLS-LDA
counts
log cpm
PLS-LDA
normal
SVM
log cpm + ash
counts shrunk

Workflow

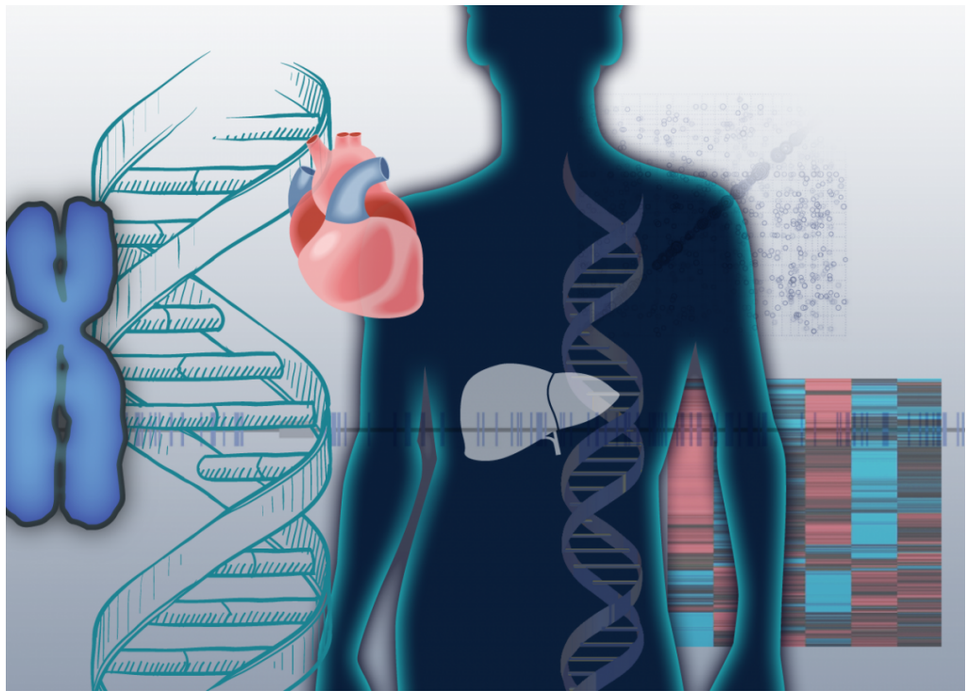


name convention:

(counts data): standard name - eg: classtpx theta.fix

(counts shrunk data) name "+ ash" (classtpx theta.fix +ash)

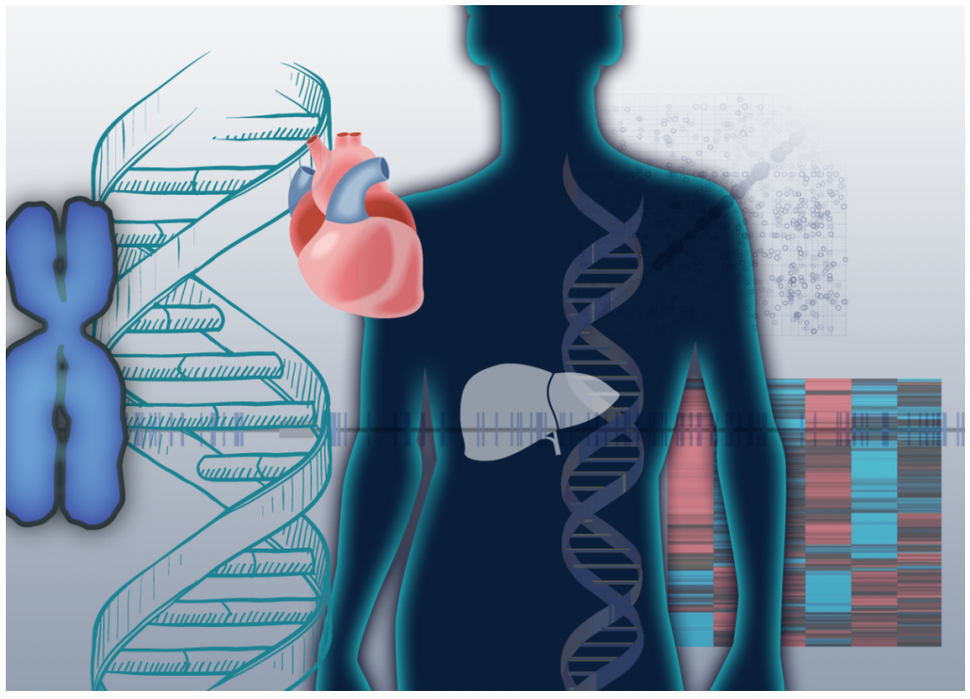
Results Summary - GTEx bulk RNA-seq



Genotype Tissue Expression Project (GTEx V6)

tissue sample RNA-seq data from 53 tissues (total 8555 samples).

We filtered out 16,000 genes initially.



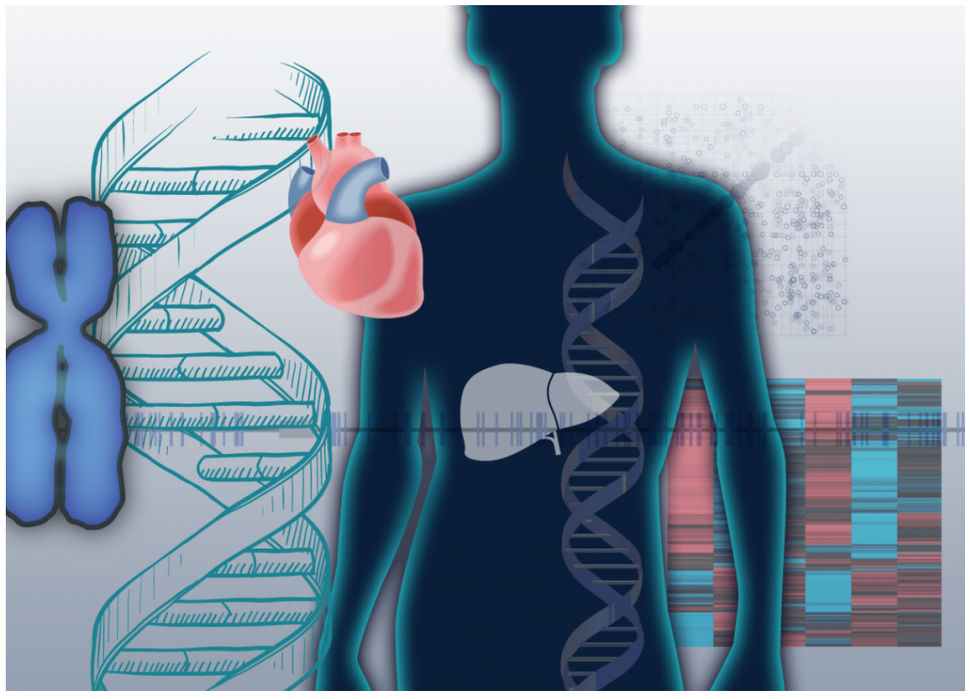
Genotype Tissue Expression Project (GTEx V6)

tissue sample RNA-seq data from 53 tissues (total 8555 samples).

We filtered out 16,000 genes initially.

For classtpx, we selected those tissues with more than 100 samples- there were 36 such tissues.

50 samples per tissue were used as training.



Genotype Tissue Expression Project (GTEx V6)

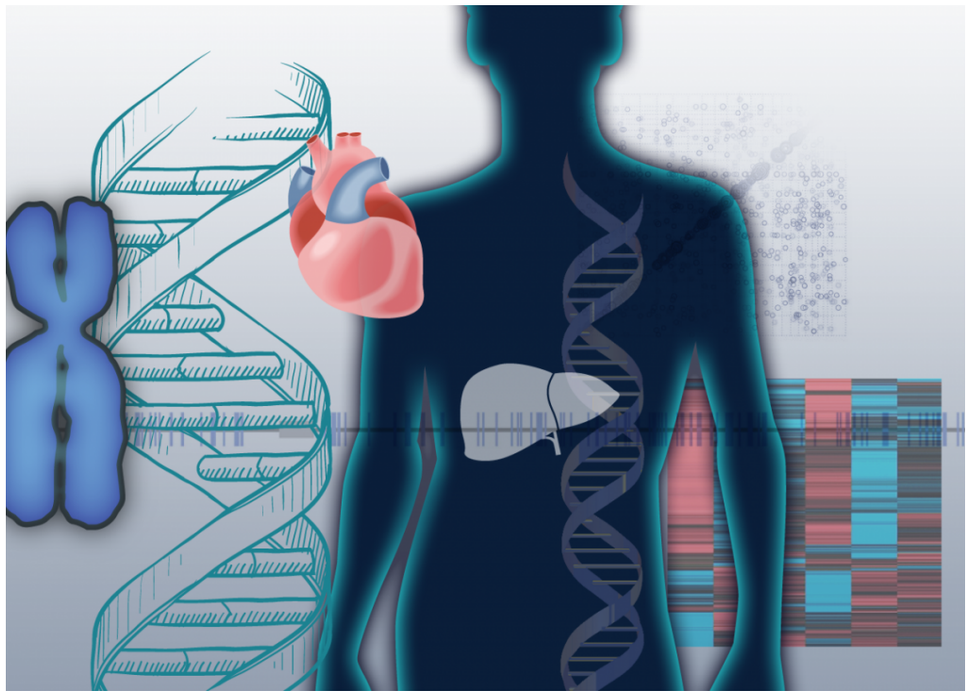
tissue sample RNA-seq data from 53 tissues (total 8555 samples).

We filtered out 16,000 genes initially.

For classtpx, we selected those tissues with more than 100 samples- there were 36 such tissues.

50 samples per tissue were used as training.

For classification on all 36 tissues, we extracted 7000 top genes distinguishing the 36 tissue classes using *CountClust::ExtractTopfeatures()* [except neural network for which top 2000 genes were used]



Genotype Tissue Expression Project (GTEx V6)

tissue sample RNA-seq data from 53 tissues (total 8555 samples).
We filtered out 16,000 genes initially.

For classtpx, we selected those tissues with more than
100 samples- there were 36 such tissues.

50 samples per tissue were used as training.

For classification on all 36 tissues, we extracted 7000 top genes
distinguishing the 36 tissue classes using *CountClust::ExtractTopfeatures()*
[except neural for which top 2000 genes were used]

Separate classification was done on pairs and triplets of tissues
- Arteries (Coronary, Tibial, Aorta),
- Breast Mammary Tissue, Adipose Subcutaneous, Adipose Visceral

For these comparison studies we did analysis on all genes and
on top 3000 genes

(a)

Brain (1259)
Cell Transformed Fibroblasts (284)
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

Whole Blood
Thyroid
Testis
Stomach
Spleen
Skin - Sun Exposed (Lower leg)
Skin - Not Sun Exposed (Suprapubic)
Prostate
Pituitary
Pancreas
Nerve - Tibial
Muscle - Skeletal
Lung
Liver
Heart - Left Ventricle
Heart - Atrial Appendage
Esophagus - Muscularis
Esophagus - Mucosa
Esophagus - Gastroesophageal Junc
Colon - Transverse
Colon - Sigmoid
Cells - Transformed fibroblasts
Cells - EBV-transformed lymphocytes
Breast - Mammary Tissue
Brain - Nucleus accumbens (basal ganglia)
Brain - Frontal Cortex (BA9)
Brain - Cortex
Brain - Cerebellum
Brain - Cerebellar Hemisphere
Brain - Caudate (basal ganglia)
Artery - Tibial
Artery - Coronary
Artery - Aorta
Adrenal Gland
Adipose - Visceral (Omentum)
Adipose - Subcutaneous

0.2 0.4 0.6 0.8 1

Classtpx (theta.fix+ash) of GTEx V6 data

Adipose - Subcutaneous
Adipose - Visceral (Omentum)
Adrenal Gland
Artery - Aorta
Artery - Coronary
Artery - Tibial
Brain - Caudate (basal ganglia)
Brain - Cerebellar Hemisphere
Brain - Cerebellum
Brain - Cortex
Brain - Frontal Cortex (BA9)
Brain - Nucleus accumbens (basal ganglia)
Breast - Mammary Tissue
Cells - EBV-transformed lymphocytes
Cells - Transformed fibroblasts
Colon - Sigmoid
Colon - Transverse
Esophagus - Gastroesophageal Junction
Esophagus - Mucosa
Esophagus - Muscularis
Heart - Atrial Appendage
Heart - Left Ventricle
Liver
Lung
Muscle - Skeletal
Nerve - Tibial
Pancreas
Pituitary
Prostate
Skin - Not Sun Exposed (Suprapubic)
Skin - Sun Exposed (Lower leg)
Spleen
Stomach
Testis
Thyroid
Whole Blood

0.0 0.2 0.4 0.6 0.8 1.0

(a)

Brain (1259)
Cell Transformed Fibroblasts (284)
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

Whole Blood
Thyroid
Testis
Stomach
Spleen
Skin - Sun Exposed (Lower leg)
Skin - Not Sun Exposed (Suprapubic)
Prostate
Pituitary
Pancreas
Nerve - Tibial
Muscle - Skeletal
Lung
Liver
Heart - Left Ventricle
Heart - Atrial Appendage
Esophagus - Muscularis
Esophagus - Mucosa
Esophagus - Gastroesophageal Junction
Colon - Transverse
Colon - Sigmoid
Cells - Transformed fibroblasts
Cells - EBV-transformed lymphocytes
Breast - Mammary Tissue
Brain - Nucleus accumbens (basal ganglia)
Brain - Frontal Cortex (BA9)
Brain - Cortex
Brain - Cerebellum
Brain - Cerebellar Hemisphere
Brain - Caudate (basal ganglia)
Artery - Tibial
Artery - Coronary
Artery - Aorta
Adrenal Gland
Adipose - Visceral (Omentum)
Adipose - Subcutaneous

0.2 0.4 0.6 0.8 1

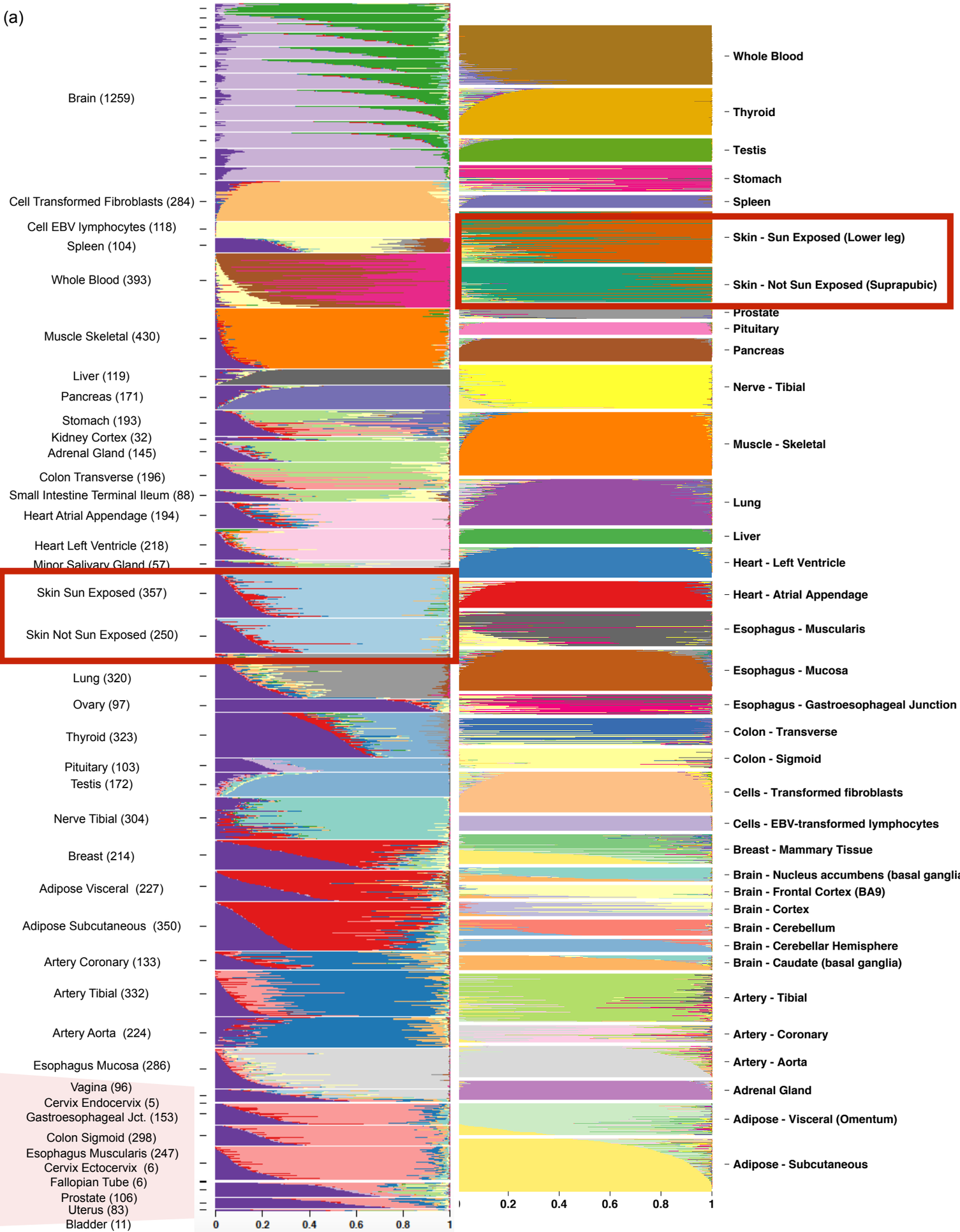
Classtpx (theta.fix+ash) of GTEx V6 data

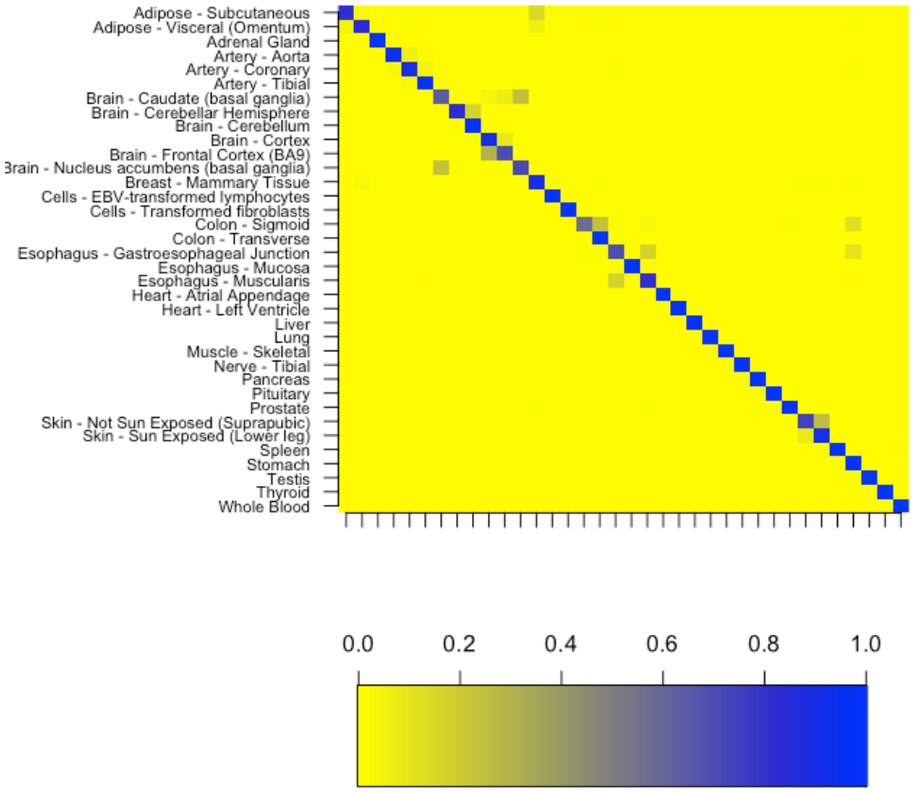
Adipose - Subcutaneous
Adipose - Visceral (Omentum)
Adrenal Gland
Artery - Aorta
Artery - Coronary
Artery - Tibial
Brain - Caudate (basal ganglia)
Brain - Cerebellar Hemisphere
Brain - Cerebellum
Brain - Cortex
Brain - Frontal Cortex (BA9)
Brain - Nucleus accumbens (basal ganglia)
Breast - Mammary Tissue
Cells - EBV-transformed lymphocytes
Cells - Transformed fibroblasts
Colon - Sigmoid
Colon - Transverse
Esophagus - Gastroesophageal Junction
Esophagus - Mucosa
Esophagus - Muscularis
Heart - Atrial Appendage
Heart - Left Ventricle
Liver
Lung
Muscle - Skeletal
Nerve - Tibial
Pancreas
Pituitary
Prostate
Skin - Not Sun Exposed (Suprapubic)
Skin - Sun Exposed (Lower leg)
Spleen
Stomach
Testis
Thyroid
Whole Blood

0.0 0.2 0.4 0.6 0.8 1.0

(a)

Classtpx (theta.fix+ash) of GTEx V6 data





(a)

Brain (1259)
Cell Transformed Fibroblasts (284)
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

Whole Blood
Thyroid
Testis
Stomach
Spleen
Skin - Sun Exposed (Lower leg)
Skin - Not Sun Exposed (Suprapubic)
Prostate
Pituitary
Pancreas
Nerve - Tibial
Muscle - Skeletal
Lung
Liver
Heart - Left Ventricle
Heart - Atrial Appendage
Esophagus - Muscularis
Esophagus - Mucosa
Esophagus - Gastroesophageal Junction
Colon - Transverse
Colon - Sigmoid
Cells - Transformed fibroblasts
Cells - EBV-transformed lymphocytes
Breast - Mammary Tissue
Brain - Nucleus accumbens (basal ganglia)
Brain - Frontal Cortex (BA9)
Brain - Cortex
Brain - Cerebellum
Brain - Cerebellar Hemisphere
Brain - Caudate (basal ganglia)
Artery - Tibial
Artery - Coronary
Artery - Aorta
Adrenal Gland
Adipose - Visceral (Omentum)
Adipose - Subcutaneous

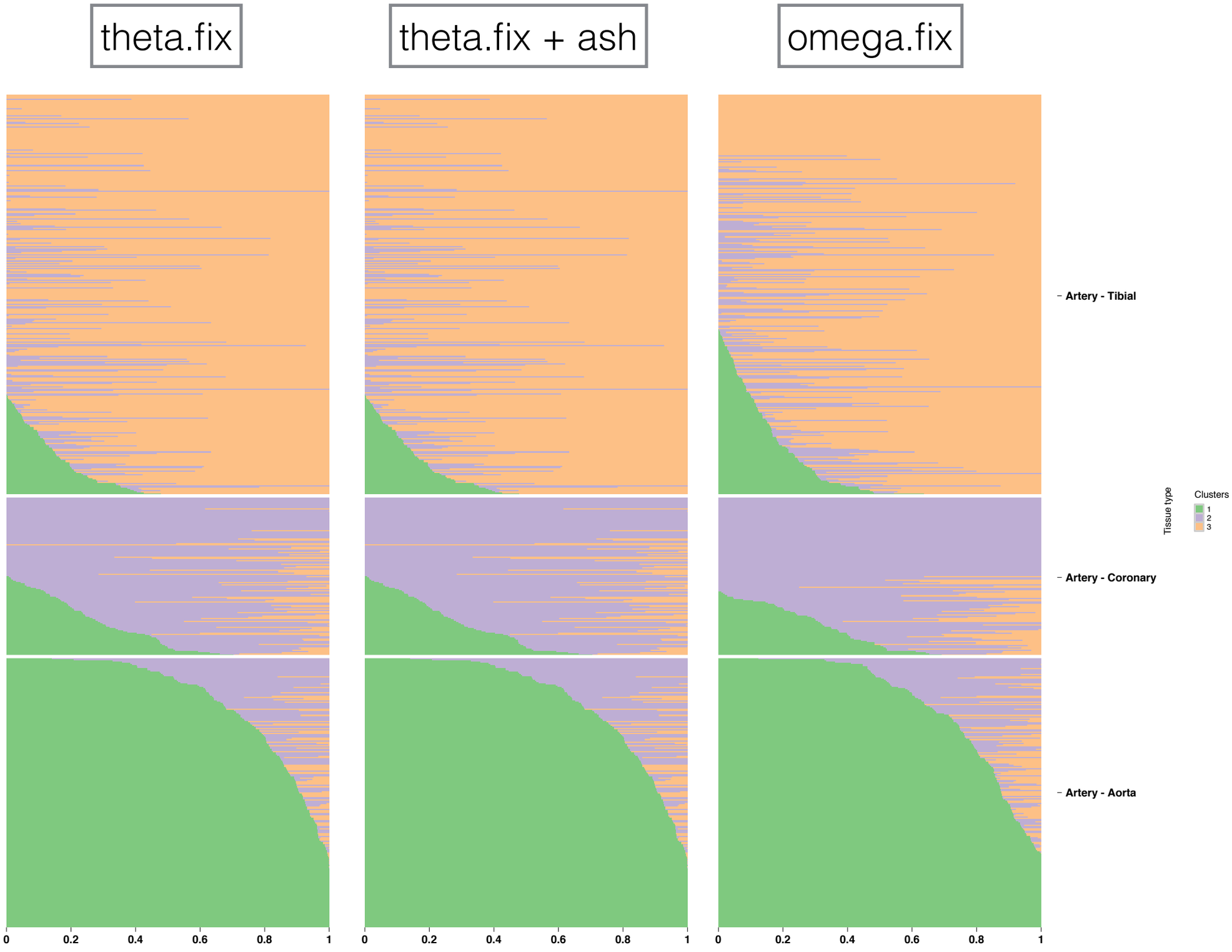
0.2 0.4 0.6 0.8 1

Classtpx (theta.fix+ash) of GTEx V6 data

Adipose - Subcutaneous
Adipose - Visceral (Omentum)
Adrenal Gland
Artery - Aorta
Artery - Coronary
Artery - Tibial
Brain - Caudate (basal ganglia)
Brain - Cerebellar Hemisphere
Brain - Cerebellum
Brain - Cortex
Brain - Frontal Cortex (BA9)
Brain - Nucleus accumbens (basal ganglia)
Breast - Mammary Tissue
Cells - EBV-transformed lymphocytes
Cells - Transformed fibroblasts
Colon - Sigmoid
Colon - Transverse
Esophagus - Gastroesophageal Junction
Esophagus - Mucosa
Esophagus - Muscularis
Heart - Atrial Appendage
Heart - Left Ventricle
Liver
Lung
Muscle - Skeletal
Nerve - Tibial
Pancreas
Pituitary
Prostate
Skin - Not Sun Exposed (Suprapubic)
Skin - Sun Exposed (Lower leg)
Spleen
Stomach
Testis
Thyroid
Whole Blood

0.0 0.2 0.4 0.6 0.8 1.0

Classtpx of three artery tissues in GTEx V6 data



classtpx vs Other classifiers

To compare with hard clustering methods like SVM, PLS-LDA etc, we assign each sample to the cluster or class that is most representative of the sample (which class has the highest grade of membership for that sample)

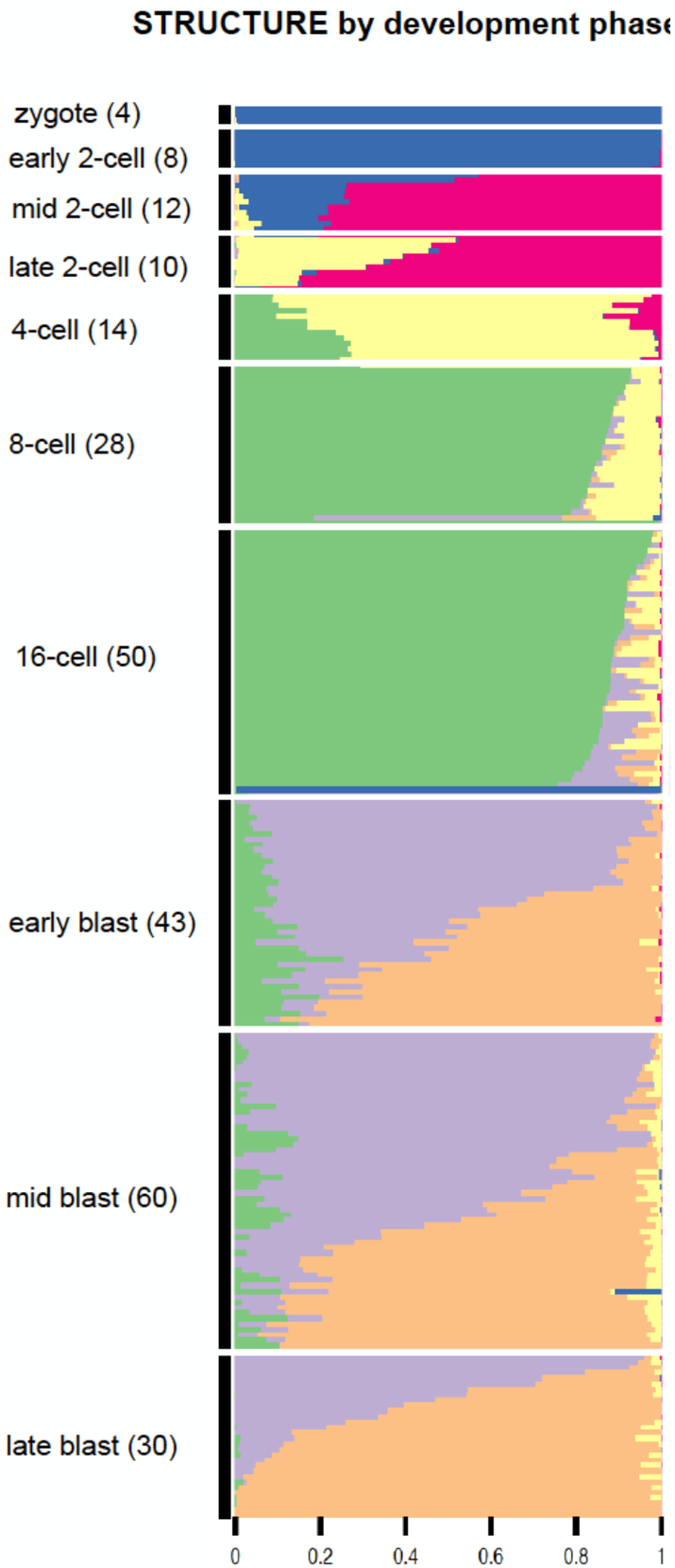
misclass prop	All tissues	Arteries (3 tissues)	Arteries3000 (3 tissues)	Breast +Adipose	Breast + Adipose 3000
classtpx theta.fix	0.07	0.09	0.08	0.14	0.13
classtpx theta.fix +ash	0.07	0.09	0.08	0.13	0.13
classtpx omega.fix	0.09	0.13	0.11	0.22	0.19
SVM (counts)	0.09	0.22	0.20	0.21	0.15
SVM (log cpm)	0.05	0.05	0.03	0.08	0.05
SVM (log cpm+ash)	0.05	0.05	0.03	0.09	0.05
PLS-LDA (counts)	0.14	0.06	0.03	0.11	0.10
PLS-LDA (log cpm)	0.03	0.22	0.2	0.24	0.06
PLS-LDA (log cpm +ash)	0.03	0.16	0.16	0.18	0.06
Normal	0.09	0.12	0.09	0.18	0.11
Poisson	0.08	0.09	0.08	0.14	0.13
Neg Binom	0.10	0.18	0.10	0.23	0.17
NN	0.21	-	-	-	-

	All tissues	Arteries (3 tissues)	Arteries3000 (3 tissues)	Breast +Adipose	Breast + Adipose 3000
classtpx theta.fix	0.07	0.09	0.08	0.14	0.13
classtpx theta.fix +ash	0.07	0.09	0.08	0.13	0.13
classtpx omega.fix	0.09	0.13	0.11	0.22	0.19
SVM (counts)	0.09	0.22	0.20	0.21	0.15
SVM (voom)	0.05	0.05	0.03	0.08	0.05
SVM (voom+ash)	0.05	0.05	0.03	0.09	0.05
PLS-LDA (counts)	0.14	0.06	0.03	0.11	0.10
PLS-LDA (voom)	0.03	0.22	0.2	0.24	0.06
PLS-LDA (voom +ash)	0.03	0.16	0.16	0.18	0.06
Normal	0.09	0.12	0.09	0.18	0.11
Poisson	0.08	0.09	0.08	0.14	0.13
Neg Binom	0.10	0.18	0.10	0.23	0.17
NN	0.21	-	-	-	-

	All tissues	Arteries (3 tissues)	Arteries3000 (3 tissues)	Breast +Adipose	Breast + Adipose 3000
classtpx theta.fix	0.07	0.09	0.08	0.14	0.13
classtpx theta.fix +ash	0.07	0.09	0.08	0.13	0.13
classtpx omega.fix	0.09	0.13	0.11	0.22	0.19
SVM (counts)	0.09	0.22	0.20	0.21	0.15
SVM (voom)	0.05	0.05	0.03	0.08	0.05
SVM (voom+ash)	0.05	0.05	0.03	0.09	0.05
PLS-LDA (counts)	0.14	0.06	0.03	0.11	0.10
PLS-LDA (voom)	0.03	0.22	0.2	0.24	0.06
PLS-LDA (voom +ash)	0.03	0.16	0.16	0.18	0.06
Normal	0.09	0.12	0.09	0.18	0.11
Poisson	0.08	0.09	0.08	0.14	0.13
Neg Binom	0.10	0.18	0.10	0.23	0.17
NN	0.21	-	-	-	-

	All tissues	Arteries (3 tissues)	Arteries3000 (3 tissues)	Breast +Adipose	Breast + Adipose 3000
classtpx theta.fix	0.07	0.09	0.08	0.14	0.13
classtpx theta.fix +ash	0.07	0.09	0.08	0.13	0.13
classtpx omega.fix	0.09	0.13	0.11	0.22	0.19
SVM (counts)	0.09	0.22	0.20	0.21	0.15
SVM (voom)	0.05	0.05	0.03	0.08	0.05
SVM (voom+ash)	0.05	0.05	0.03	0.09	0.05
PLS-LDA (counts)	0.14	0.06	0.03	0.11	0.10
PLS-LDA (voom)	0.03	0.22	0.2	0.24	0.06
PLS-LDA (voom +ash)	0.03	0.16	0.16	0.18	0.06
Normal	0.09	0.12	0.09	0.18	0.11
Poisson	0.08	0.09	0.08	0.14	0.13
Neg Binom	0.10	0.18	0.10	0.23	0.17
NN	0.21	-	-	-	-

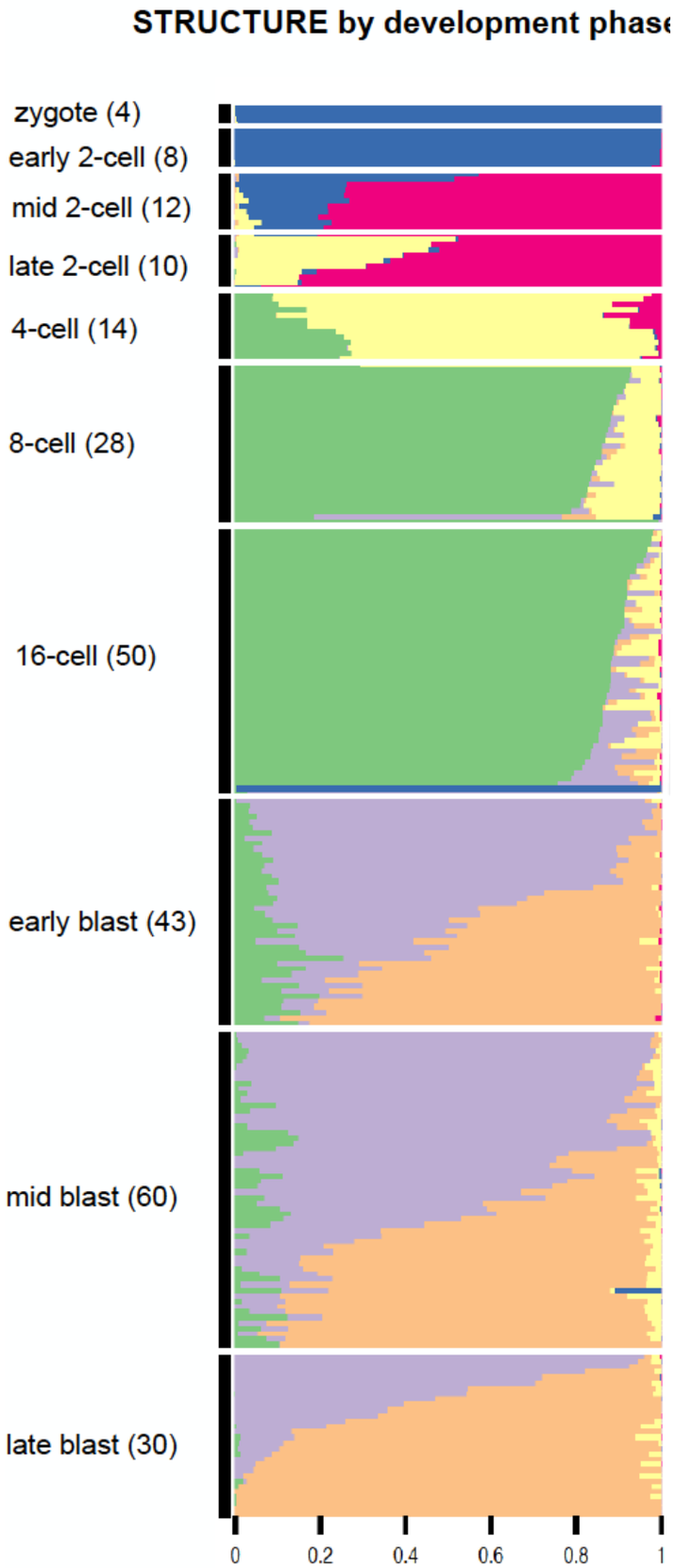
How about single-cell data?



6 clusters

Deng et al (2014) single cell development phase data

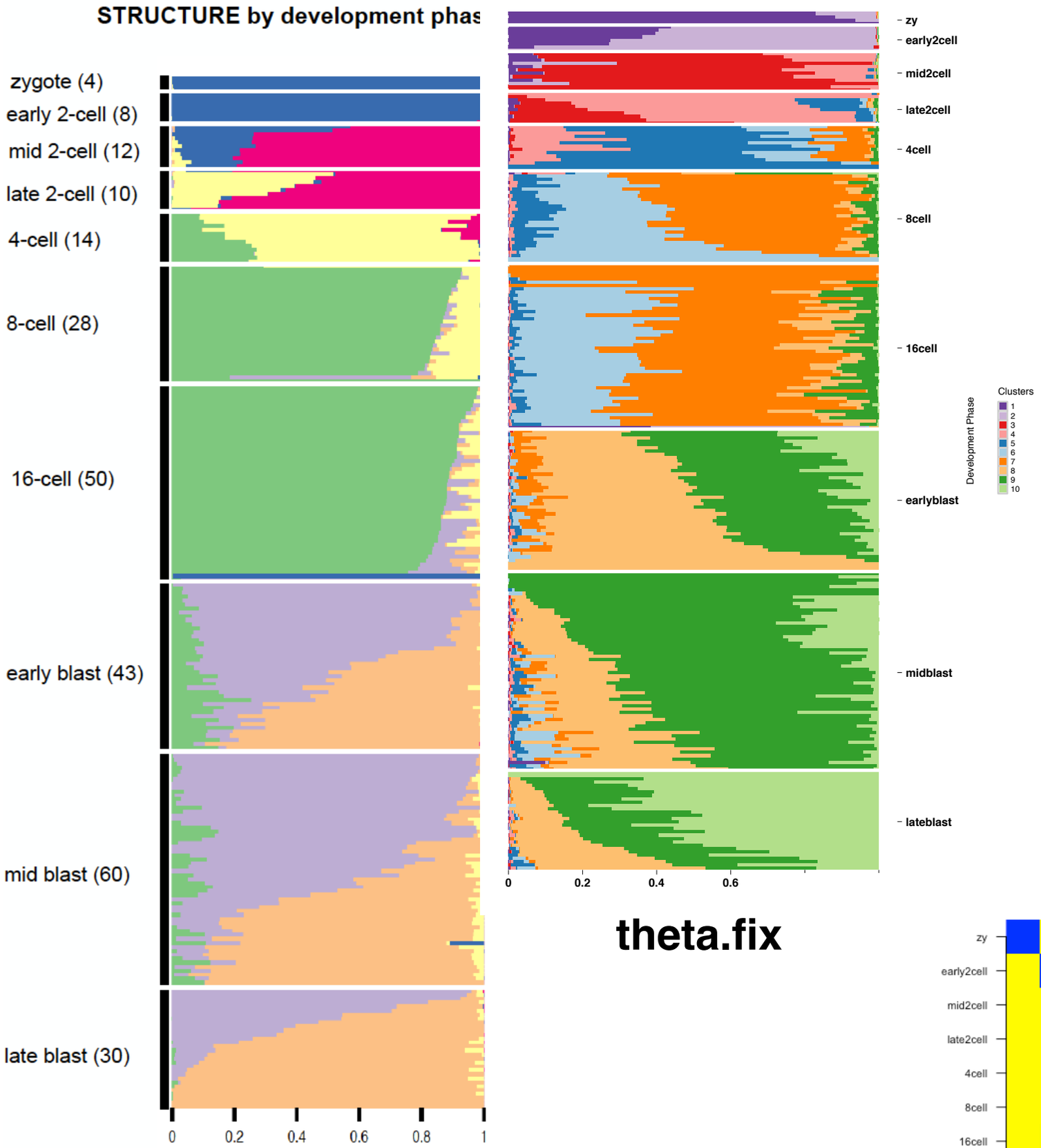
maptpx clustering



6 clusters

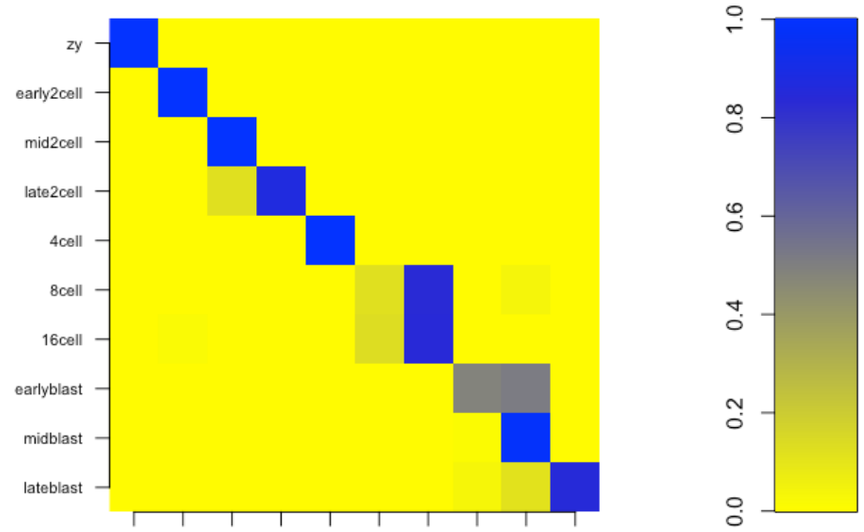
Deng et al (2014) single cell development phase data

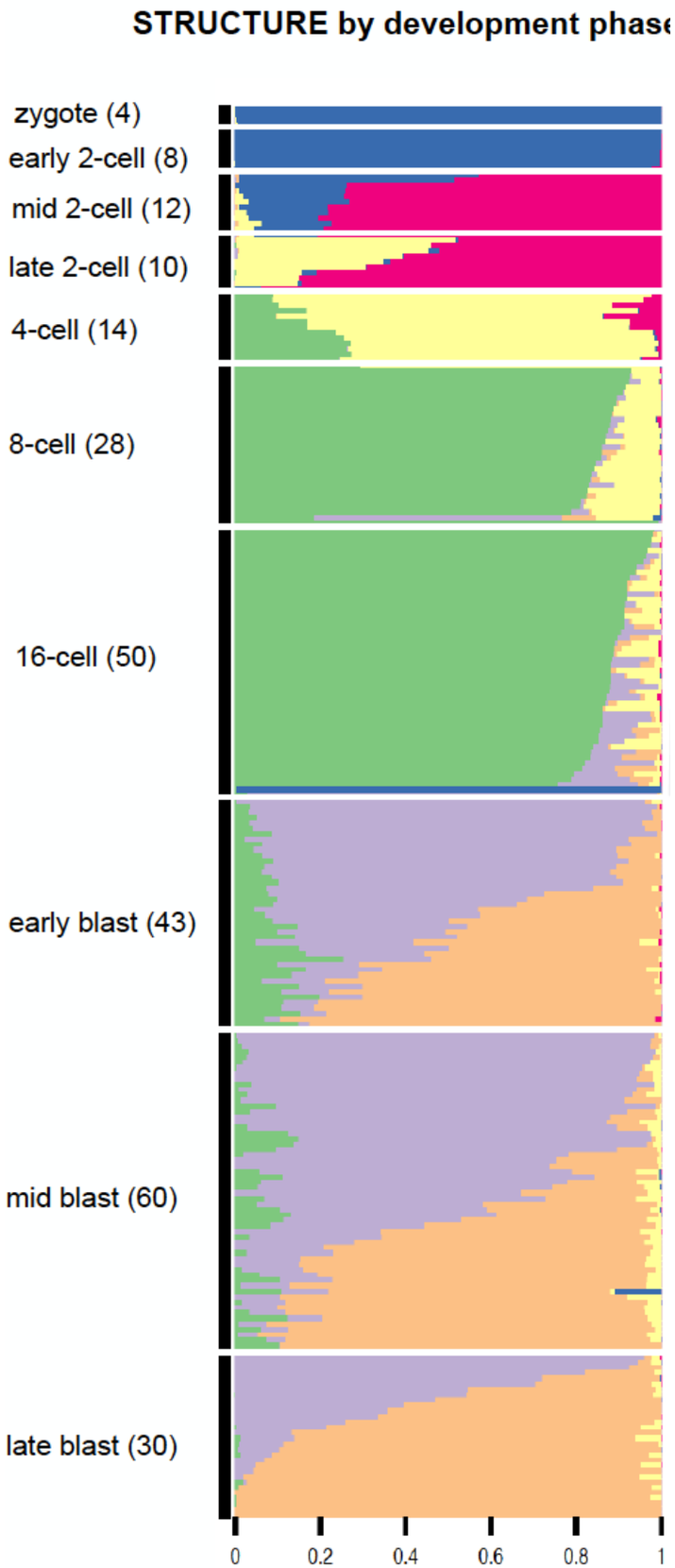
We pick 20 % samples from each phase as training for classtpx and other classifiers



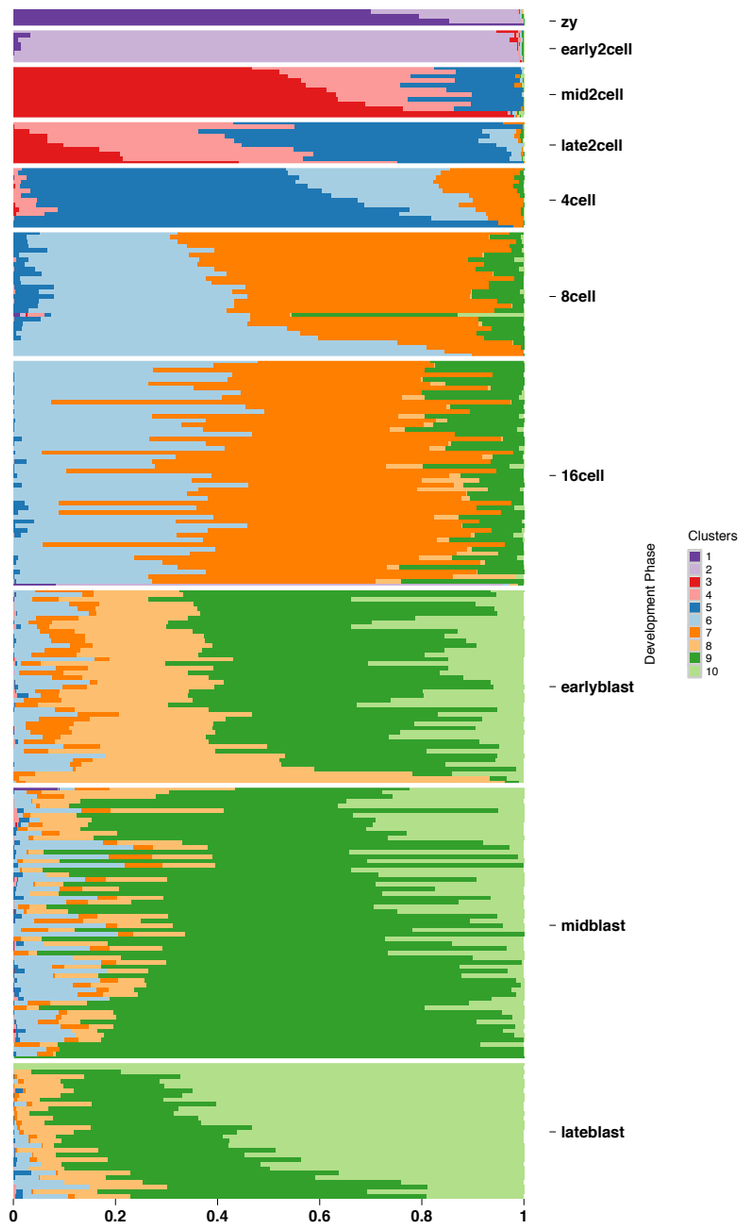
6 clusters

theta.fix

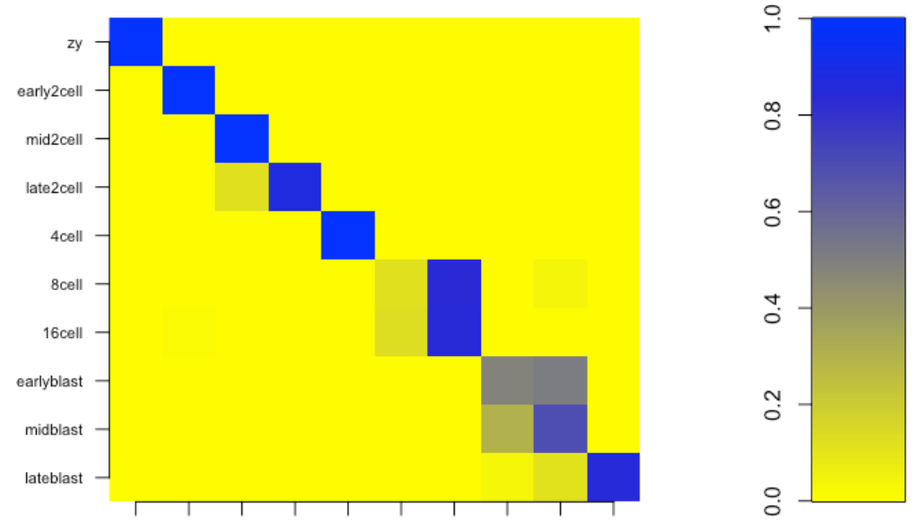


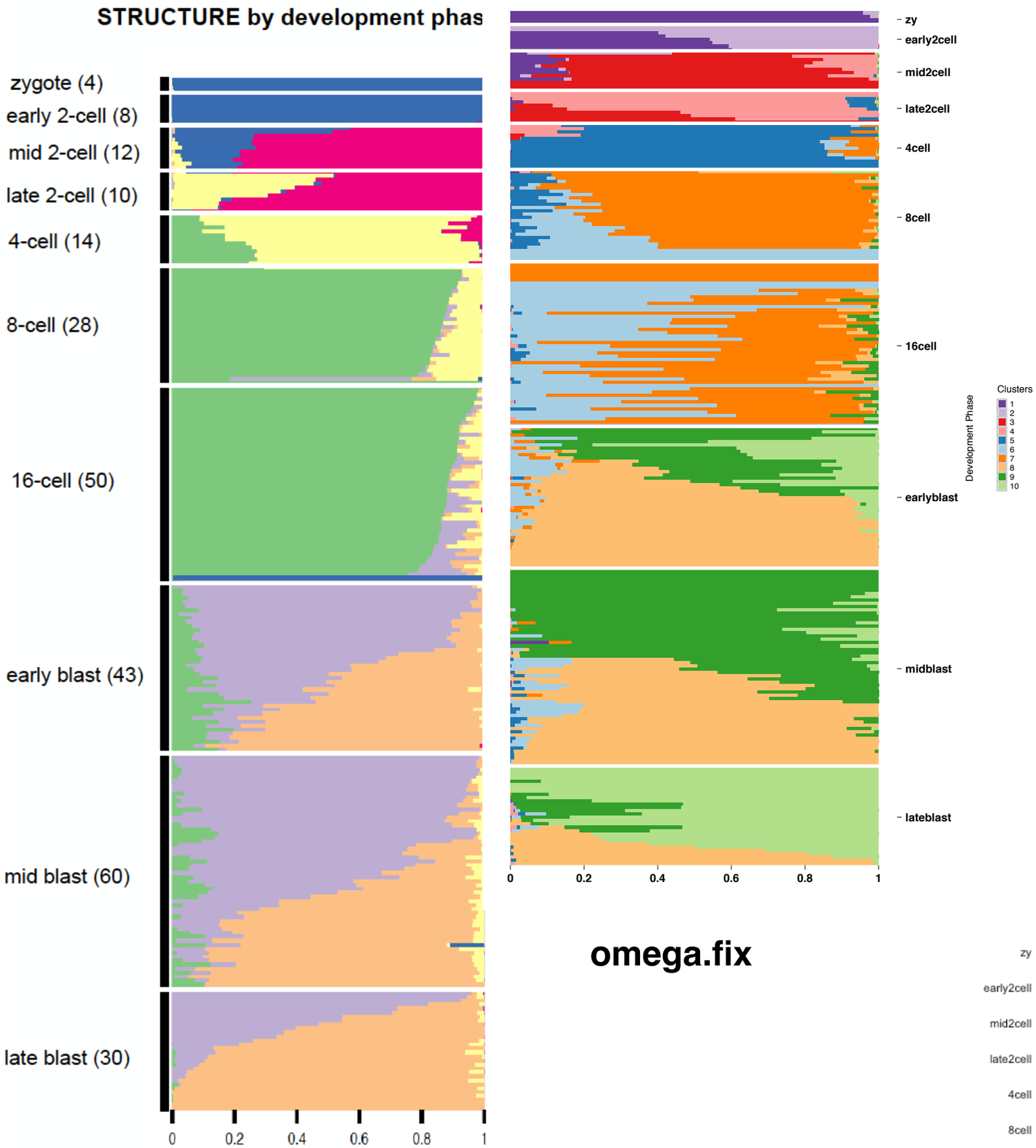


6 clusters



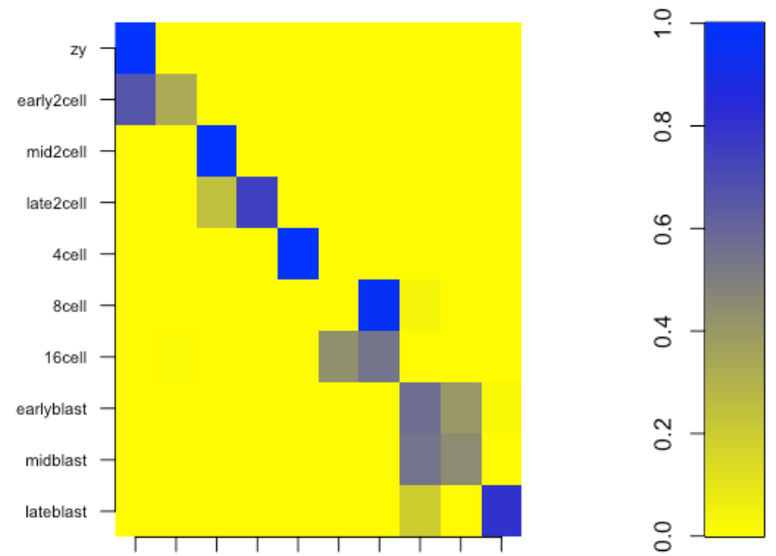
theta.fix+ash





6 clusters

omega.fix



	Deng et al (2014)
classtpx theta.fix	0.24
classtpx theta.fix +ash	0.31
classtpx omega.fix	0.40
SVM (counts)	> 0.6
SVM (voom)	0.56
SVM (voom+ash)	> 0.6
PLS-LDA (counts)	0.43
PLS-LDA (voom)	0.38
PLS-LDA (voom+ash)	0.59
Normal	> 0.6
Poisson	0.32
Neg Binom	0.30

Summary

classtpx seems to be robust to the size of the training sample, unlike other classifiers (SVM, PLS-LDA etc) as in Deng et al (2014)

theta.fix and theta.fix+ash seem to perform better as classier than omega.fix

As a classifier, SVM on voom or voom+ash data seems to outperform classtpx, but not by a big margin

classtpx performs soft clustering- assigns grade of membership of test data to each class, other methods not meant to serve that purpose.