Changing mixtures does not always change margins: an application to single-cell RNA-Seq

Michael A. Newton

JSM 2018, Vancouver

CPCP

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

30,000 ft view…

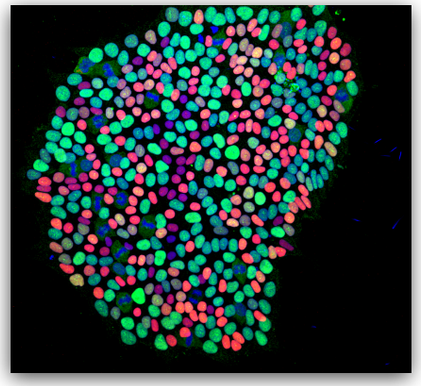
clustering

hypothesis testing

large-scale inference

Outline

- Single-cell RNA-Seq and *differential distribution*

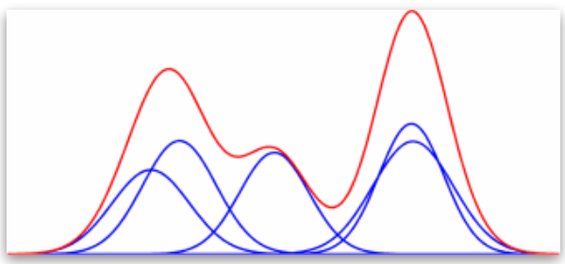


- empirical operating characteristics of a new method

- new method

scDDboost

- Empirical Bayes: *Considering two mixtures over the same same finite set of possibly non-distinct components, we report a formula for the posterior probability that the marginal mixtures are equal.*

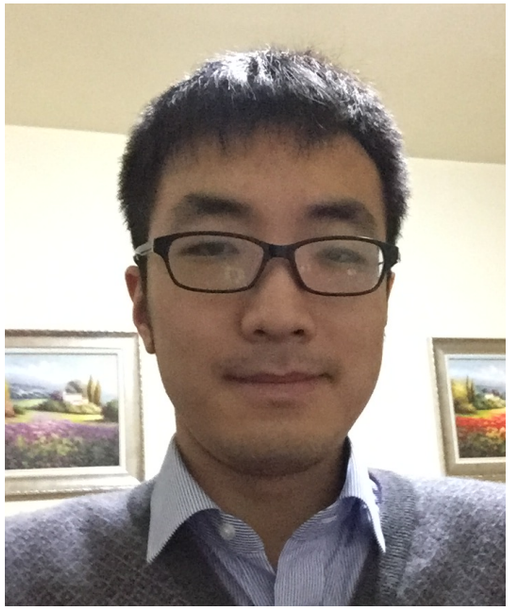


More and more scientists are jumping into single-cell analysis…and as the technologies to study single cells expand, they will require sophisticated analytical tools to tame and make sense of the results

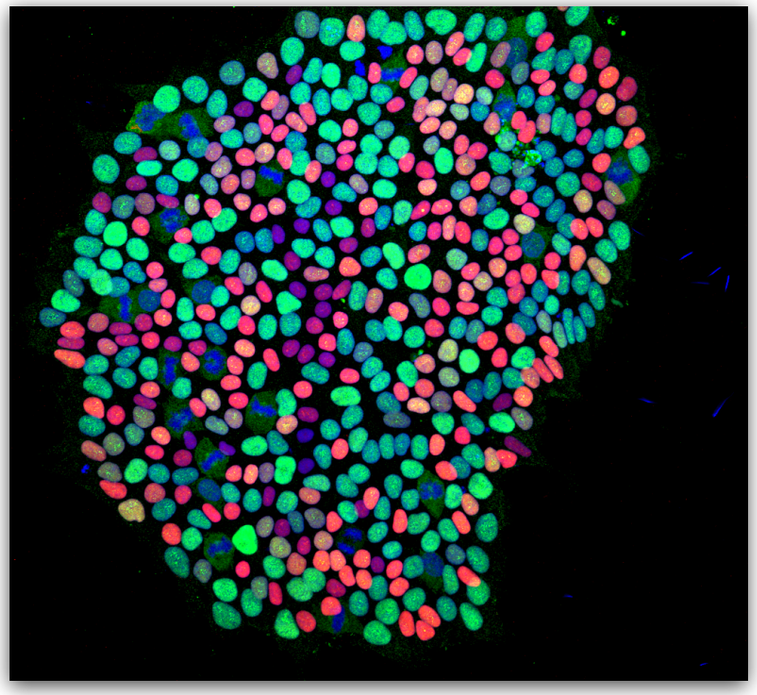
News Feature, *Nature* 547 (19): July 2017

A COMPOSITIONAL MODEL TO ASSESS EXPRESSION CHANGES FROM SINGLE-CELL RNA-SEQ DATA

XIUYU MA, CHRISTINA KENDZIORSKI, AND MICHAEL A. NEWTON



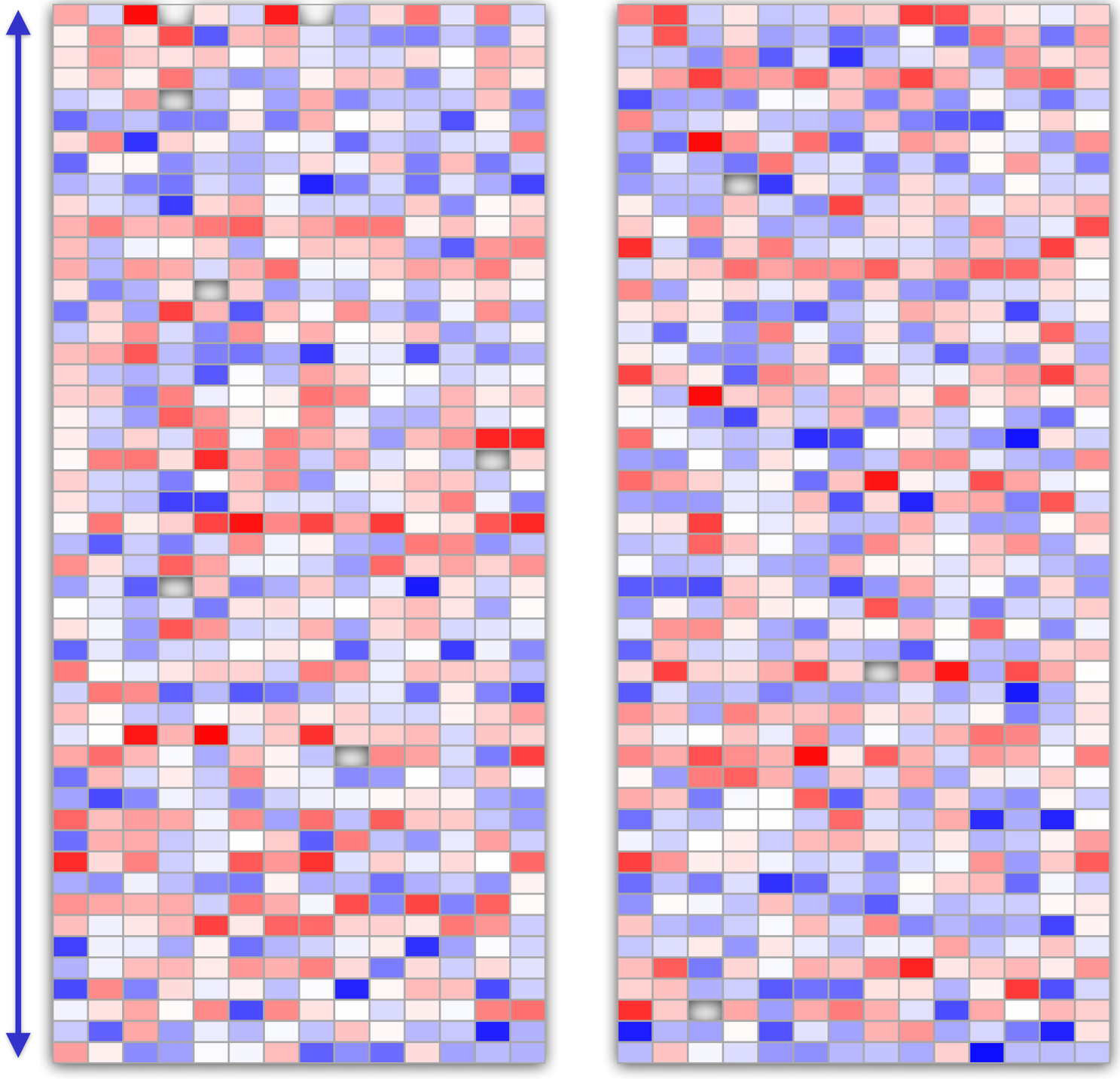
Xiuyu Ma



Data structure

cells

genes



scRNA-seq

condition 1

condition 2

Inference task

- *score genes for changes between conditions in the distribution of gene expression*

differential expression

differential distribution

Inference task

- *score genes for changes between conditions in the distribution of gene expression*

differential expression

differential distribution

- *improve sensitivity*

Inference task

- *score genes for changes between conditions in the distribution of gene expression*

differential expression

differential distribution

- *improve sensitivity [novel strength-borrowing approach]*

Some current options

- bulk methods: DESeq2 EBSeq edgeR …

Some current options

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

- bulk methods: DESeq2 EBSeq edgeR …

  - *but, variation characteristics of single-cell data are different from bulk RNA-seq data*

REVIEW Open Access

Design and computational analysis of single-cell RNA-sequencing experiments

Rhonda Bacher[1] and Christina Kendziorski[2*]

Genome Biology (2016) 17:63

METHOD Open Access

MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data

Greg Finak[1†], Andrew McDavid[1†], Masanao Yajima[1†], Jingyuan Deng[1], Vivian Gersuk[2], Alex K. Shalek[3,4,5,6], Chloe K. Slichter[1], Hannah W. Miller[1], M. Juliana McElrath[1], Martin Prlic[1], Peter S. Linsley[2] and Raphael Gottardo[1,7*]

Genome Biology (2015) 16:278

• • •

Some current options

- bulk methods: DESeq2 EBSeq edgeR …
- Single-cell differential expression scde

Kharchenko *Nature Methods* **11**, 740–742 (2014)

- Model-based analysis of single-cell transcripts MAST

Finak *et al. Genome Biology* (2015) 16:278

- Single-cell differential distribution scDD

Korthauer *et al. Genome Biology* (2016) 17:222

…

Empirical operating characteristics

scDDboost

Synthetic data:


METHOD

Splatter: simulation of single-cell RNA sequencing data

Zappia *et al. Genome Biology* (2017) 18:174

- 200 cells/condition
- Each condition a mixture of 7 distinct cell sub-populations
- details on DE between sub-population and mixture rates

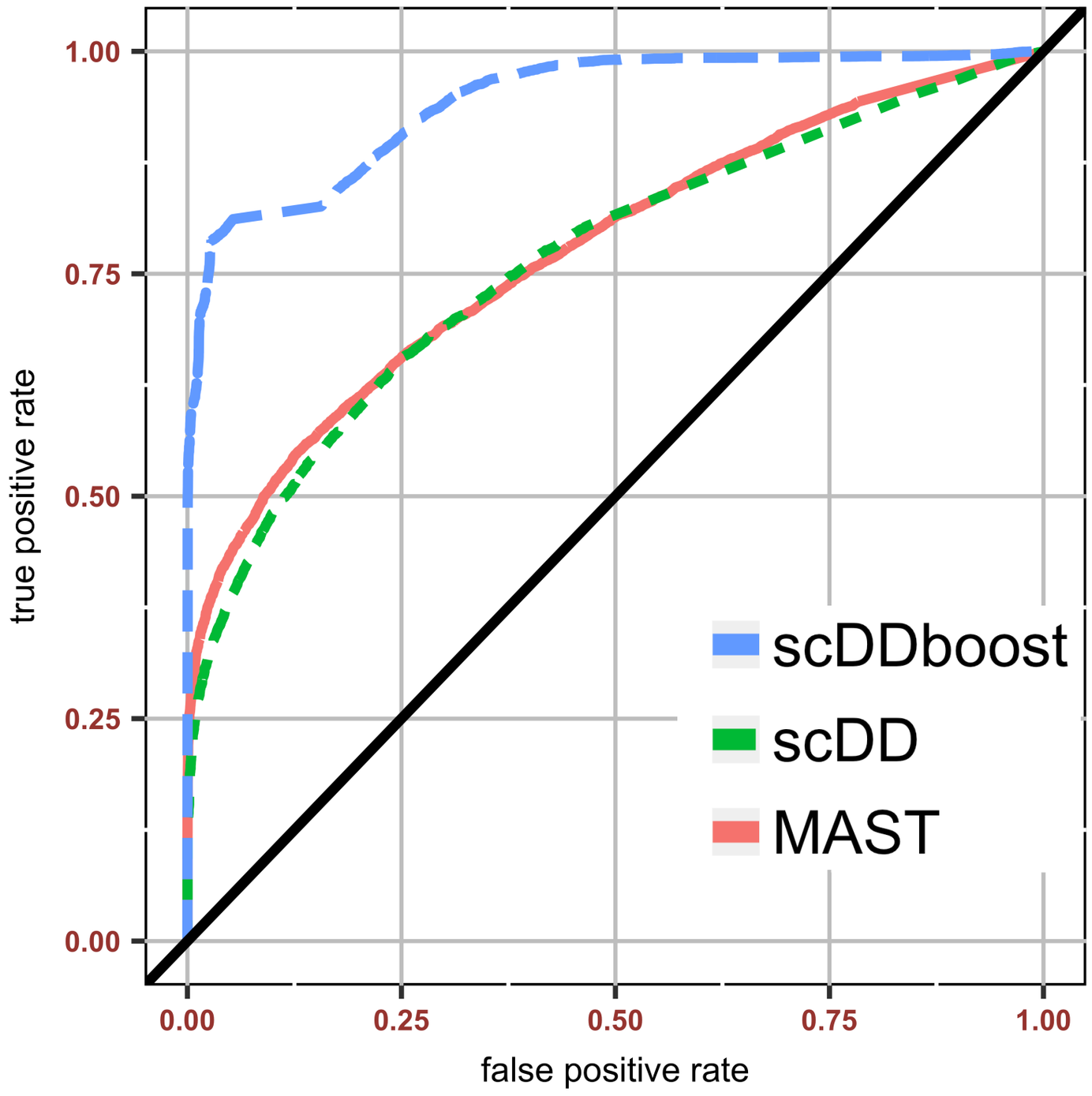
Synthetic data:

- 200 cells/condition
- Each condition a mixture of 7 distinct cell sub-populations
- details on DE between sub-population and mixture rates

METHOD

Splatter: simulation of single-cell RNA sequencing data

Zappia *et al. Genome Biology* (2017) 18:174



Empirical study

conquer (**con**sistent **qu**antification of **e**xternal **r**na-seq data)

C Soneson & MD Robinson: Bias, robustness and scalability in single-cell differential expression analysis. Nature Methods 15(4):255-261 (2018).

Empirical study

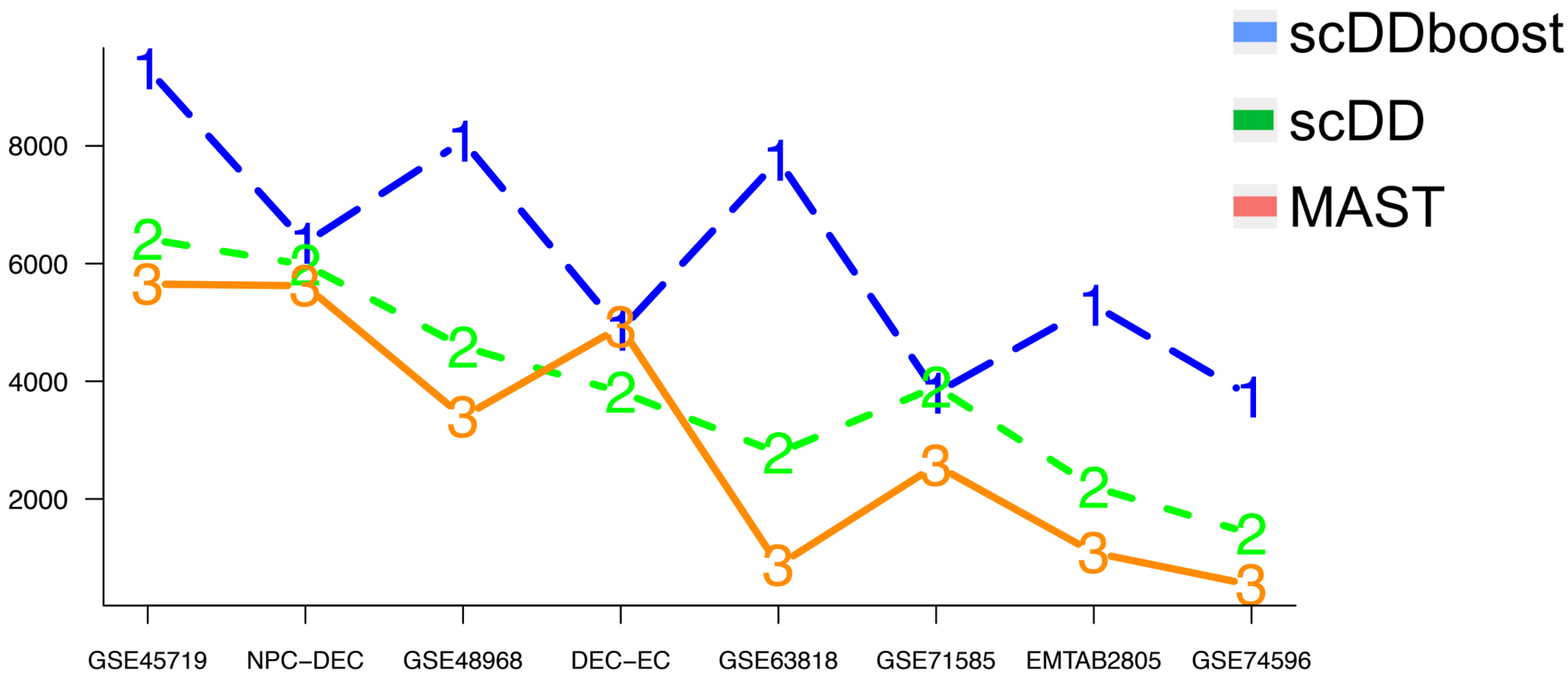
conquer (**con**sistent **qu**antification of **e**xternal **r**na-seq data)

C Soneson & MD Robinson: Bias, robustness and scalability in single-cell differential expression analysis. Nature Methods 15(4):255-261 (2018).

Data set	Conditions	Number of cells/condition	Organism
GSE74596	NKT0 vs NKT17	45,44	mouse
GSE63818-GPL16791	Primordial Germ Cells, developmental stage: 7 week gestation vs Somatic Cells, developmental stage: 7 week gestation	39,26	mouse
GSE48968-GPL13112	BMDC (1h LPS stimulation) vs BMDC(4h LPS stimulation)	96, 95	mouse
GSE45719	16-cell stage blastomere vs Mid blastocyst cell (92-94h post- fertilization)	50, 60	mouse
EMTAB2805	G1 vs G2M	96,96	mouse
GSE71585-GPL13112	Chrna2 tdTpositive vs Cux2 tdTpositive	84, 124	mouse
GSE75748	DEC vs EC	70, 64	human
GSE75748	NPC vs DEC	64, 87	human

Empirical study

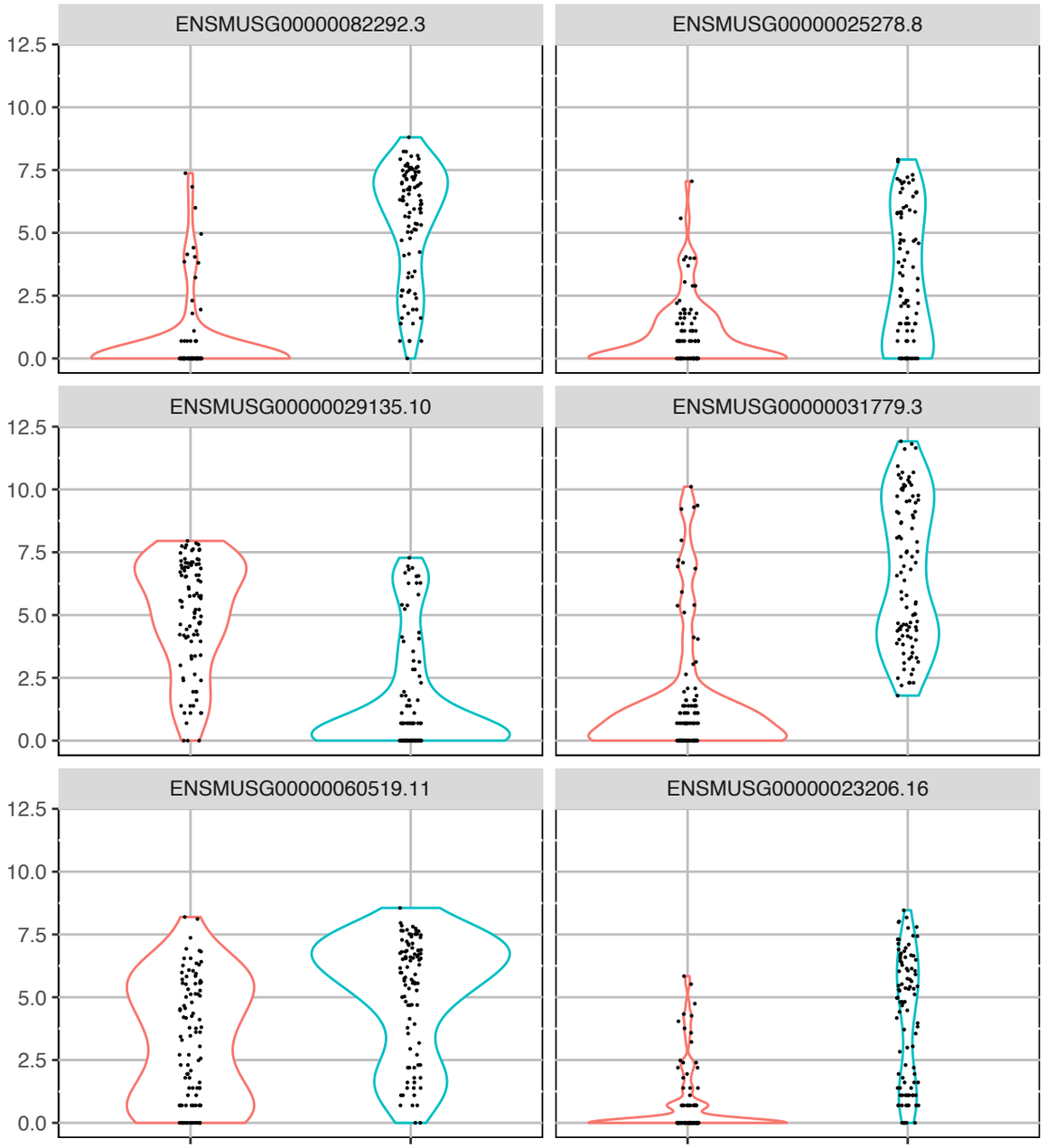
Size of 5% FDR list

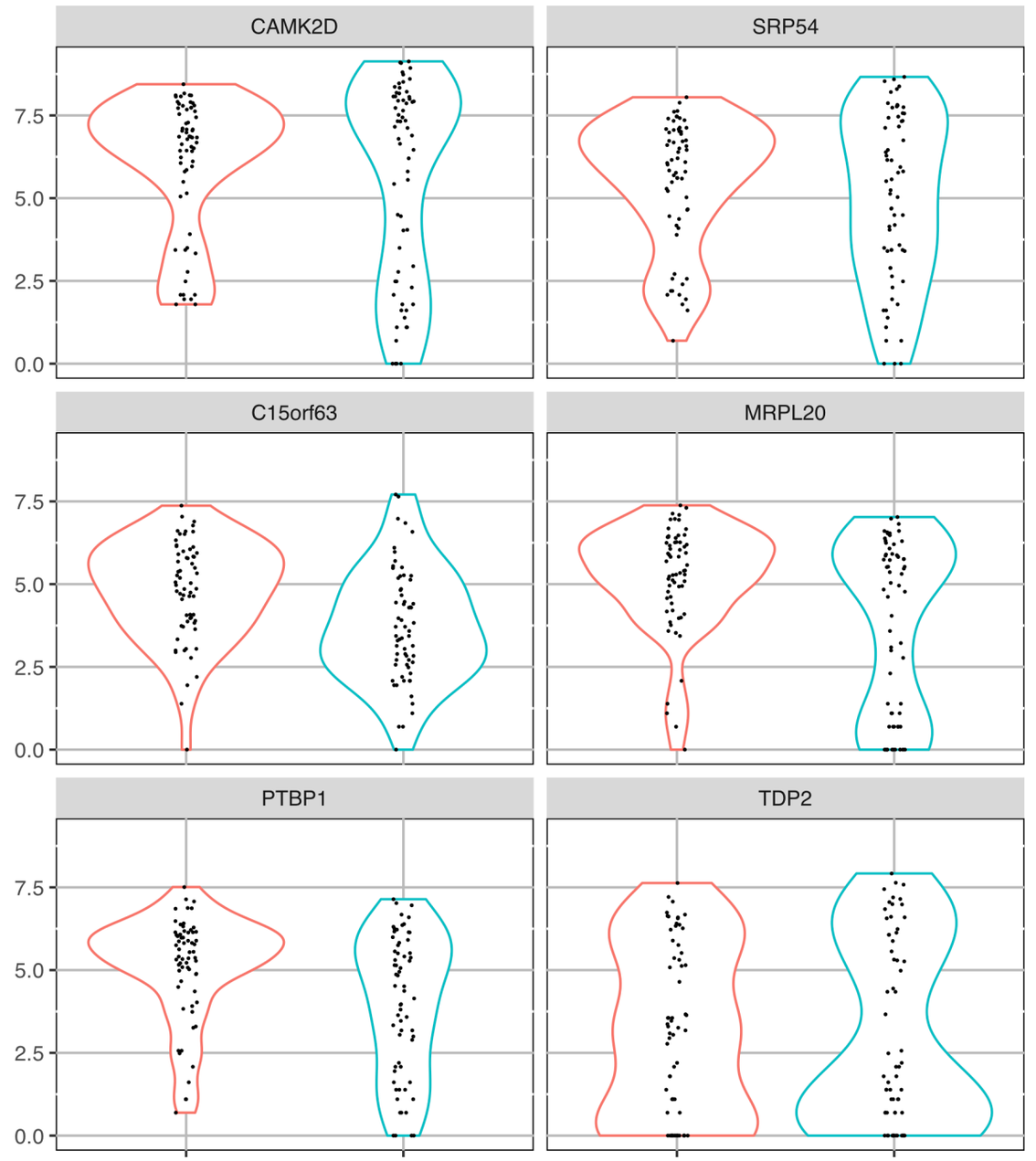


data set

ranked by mean list size

Example findings





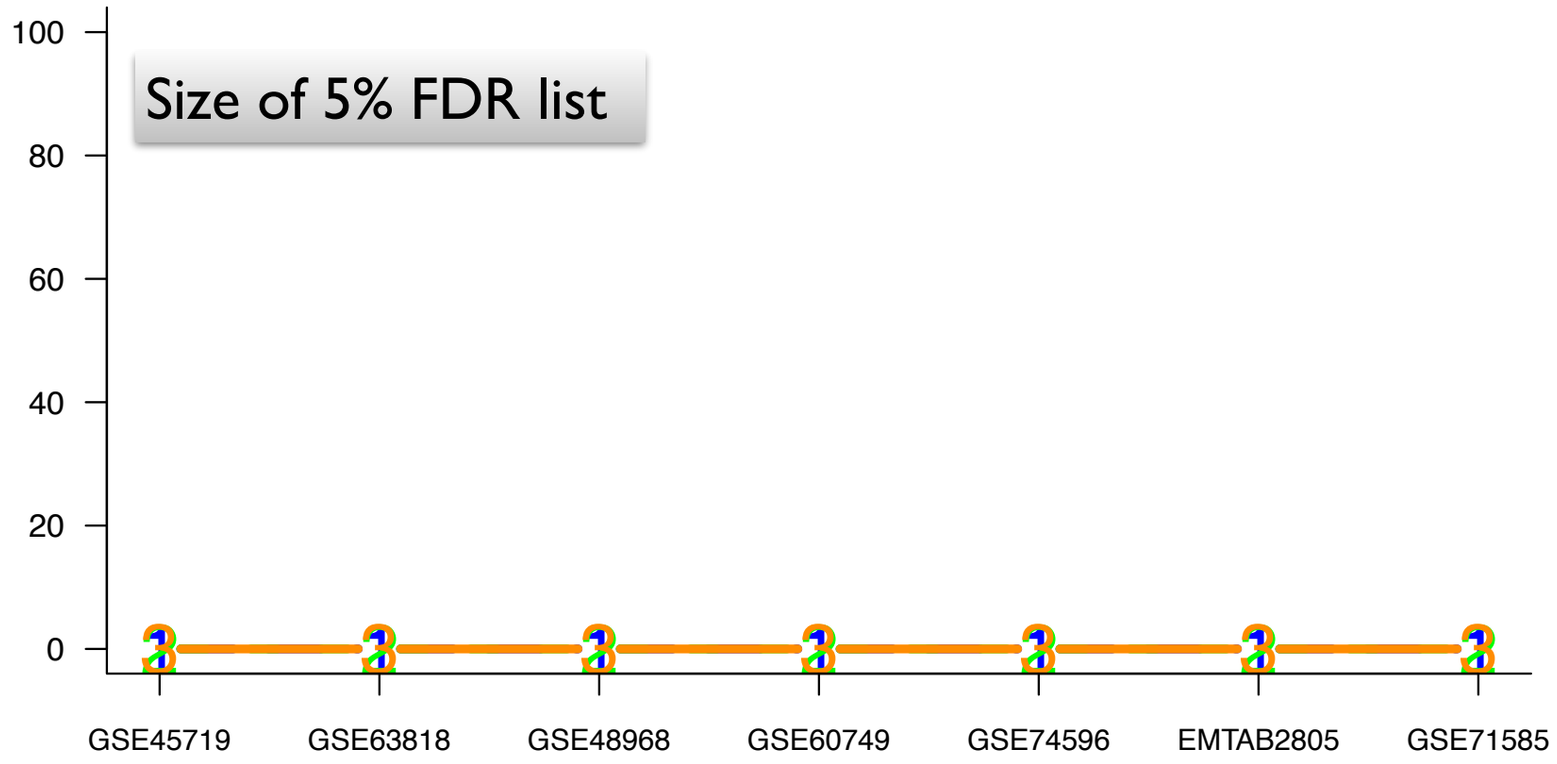
all methods

only scDDboost

Empirical study: null examples

Data set	Conditions	Number of cells/condition	Organism
GSE45719null	16-cell stage blastomere	25,25	mouse
GSE63818null	7 week gestation	20,19	mouse
GSE48968-GPL13112null	BMDC (1h LPS stimulation)	48,48	mouse
GSE60749-GPL13112null	v6.5 mouse embryonic stem cells, culture conditions: 2i+LIF	45,45	mouse
GSE74596null	NKT0	23,22	mouse
EMTAB2805null	G1	48,48	mouse
GSE71585-GPL13112null	Chrna2 tdTpositive	42,42	mouse

All three methods report no genes at 5% FDR



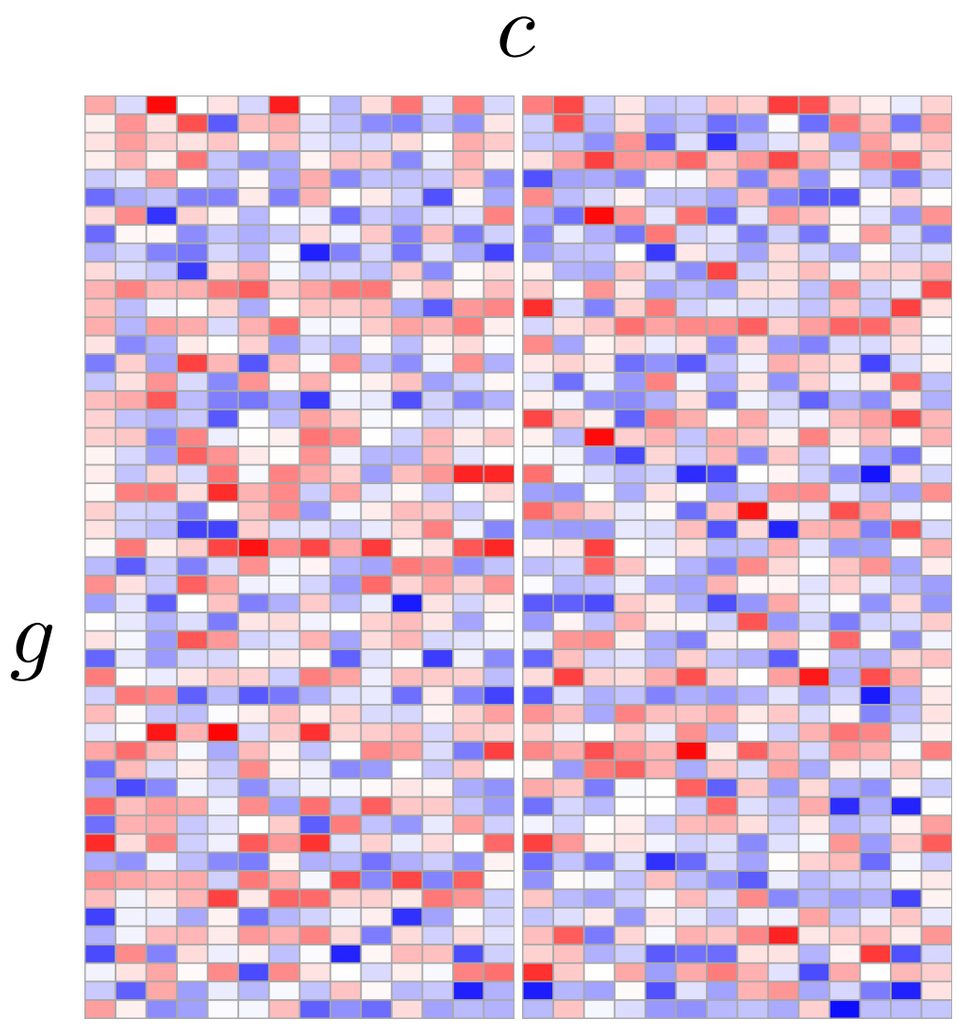
Methodology

- compositional model
- Empirical Bayes testing
- cell clustering

Notation

$X_{g,c} =$ expression of gene g in cell c

$X = (X_{g,c})$



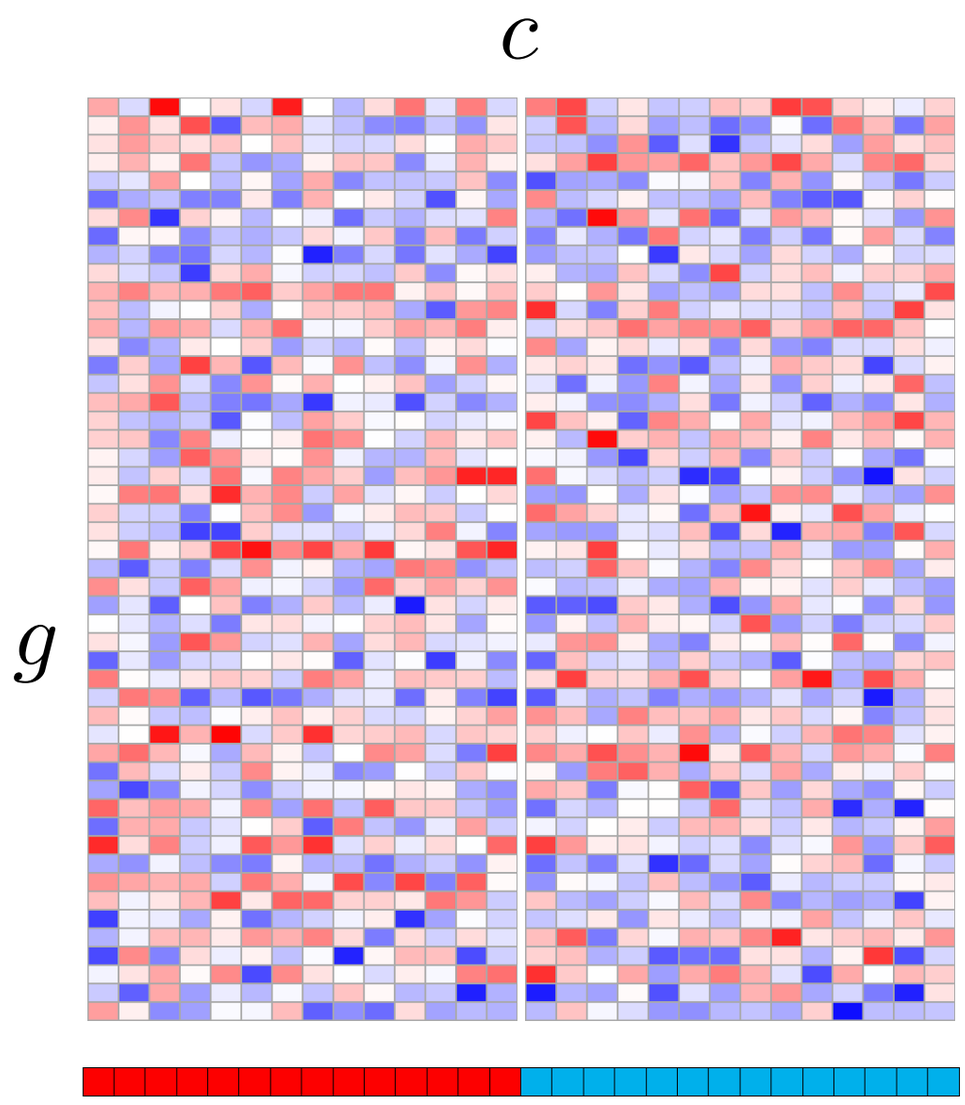
Notation

$X_{g,c} =$ expression of gene g in cell c

$X = (X_{g,c})$

$y_c =$ condition label, cell c

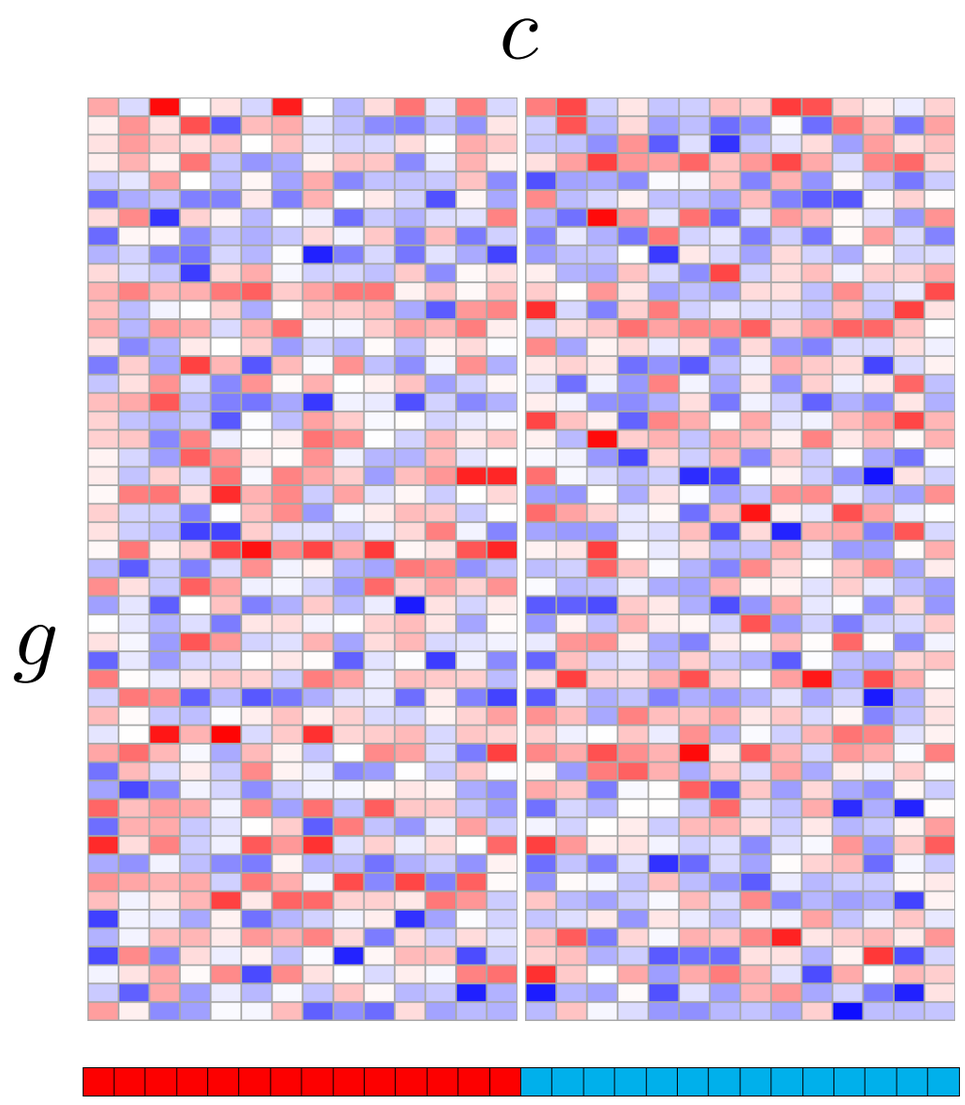
$y = (y_c)$



Notation

$X_{g,c} =$ expression of gene g in cell c

$X = (X_{g,c})$

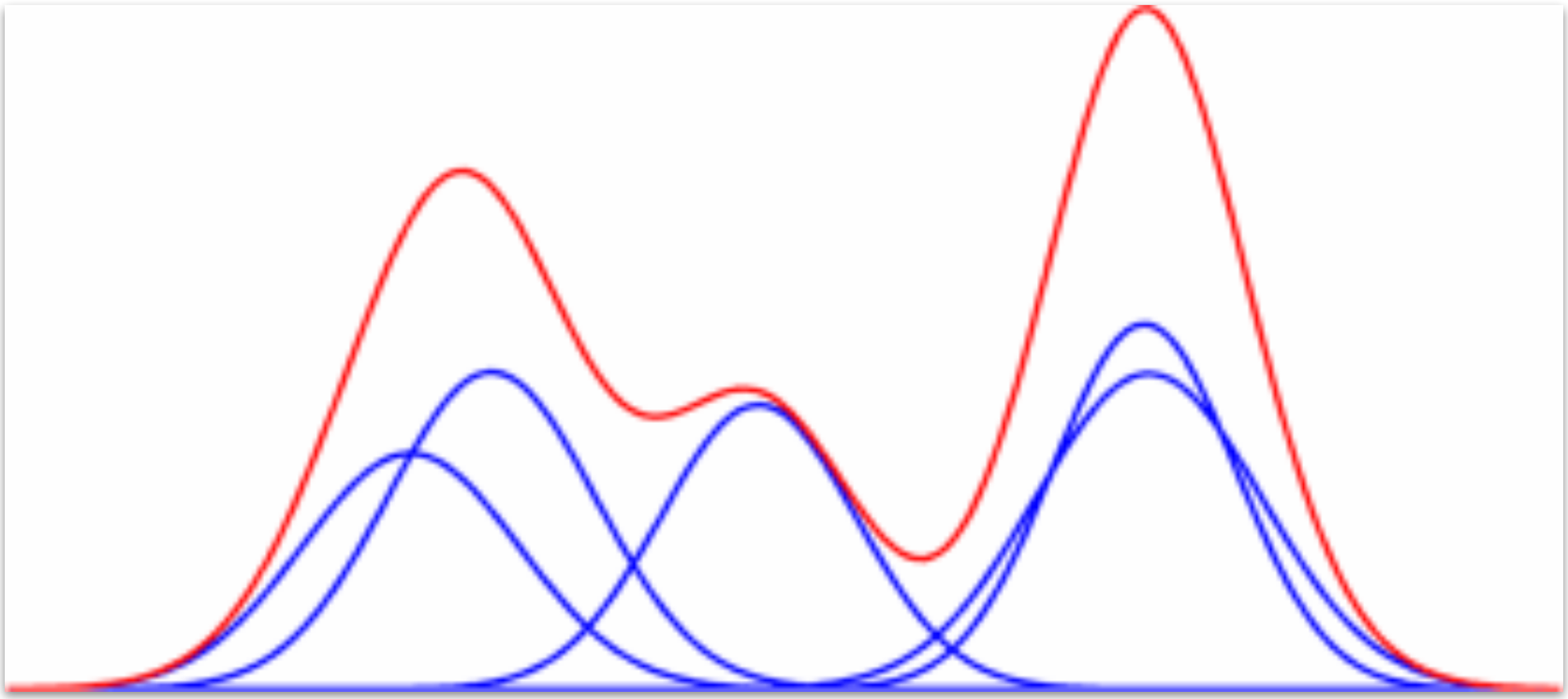


$y_c =$ condition label, cell c

$y = (y_c)$

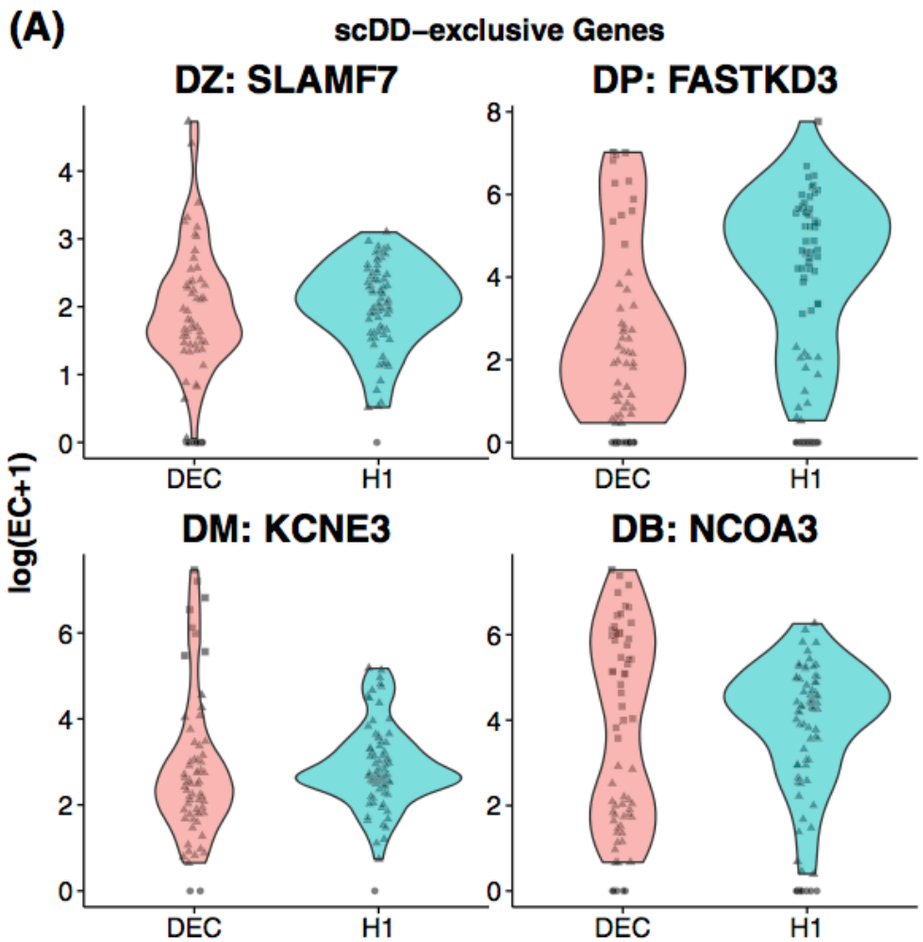
Data: (X, y)

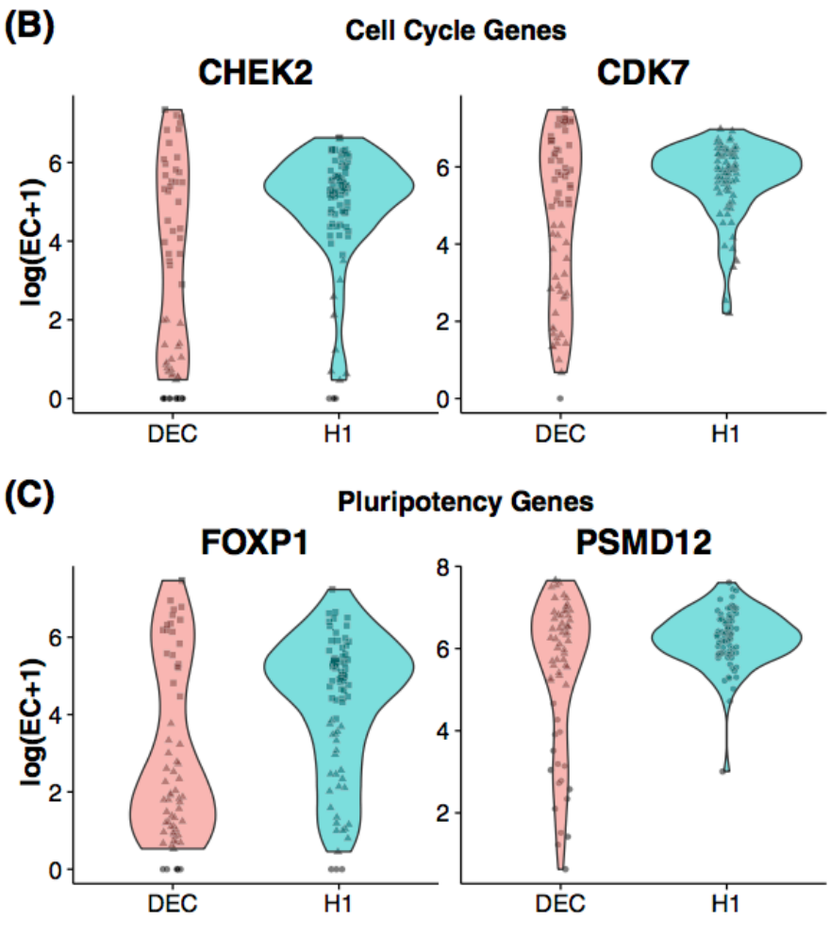
Finite mixtures



Finite mixtures

scDD and other model-based approaches assume *finite mixture* distributions per gene





H1 vs DEC

Finite mixtures

At gene g

For a cell c in one condition, say $y_c = 1$,

$$X_{g,c} \sim f_g^1(x) = \sum_{k=1}^{K} \phi_k f_{g,k}(x)$$

Finite mixtures

At gene g

For a cell c in one condition, say $y_c = 1$,

of components

$$X_{g,c} \sim f_g^1(x) = \sum_{k=1}^{K} \phi_k f_{g,k}(x)$$

Finite mixtures

At gene g

For a cell c in one condition, say $y_c = 1$,

# of components

component distributions

$$X_{g,c} \sim f_g^1(x) = \sum_{k=1}^{K} \phi_k f_{g,k}(x)$$

Finite mixtures

At gene g

For a cell c in one condition, say $y_c = 1$,

# of components

component distributions

$$X_{g,c} \sim f_g^1(x) = \sum_{k=1}^{K} \phi_k f_{g,k}(x)$$

mixing proportions

Compositional model

At gene g

For a cell c in condition $y_c = 1$

$$f_g^1(x) = \sum_{k=1}^{K} \phi_k f_{g,k}(x)$$

For a cell c in condition $y_c = 2$

$$f_g^2(x) = \sum_{k=1}^{K} \psi_k f_{g,k}(x)$$

Compositional model

At gene g

For a cell c in condition $y_c = 1$

$$f_g^1(x) = \sum_{k=1}^{K} \phi_k f_{g,k}(x)$$

same components

For a cell c in condition $y_c = 2$

$$f_g^2(x) = \sum_{k=1}^{K} \psi_k f_{g,k}(x)$$

Compositional model

At gene g

For a cell c in condition $y_c = 1$

$$f_g^1(x) = \sum_{k=1}^{K} \phi_k f_{g,k}(x)$$

For a cell c in condition $y_c = 2$

$$f_g^2(x) = \sum_{k=1}^{K} \psi_k f_{g,k}(x)$$

possibly different mixing rates

Compositional model

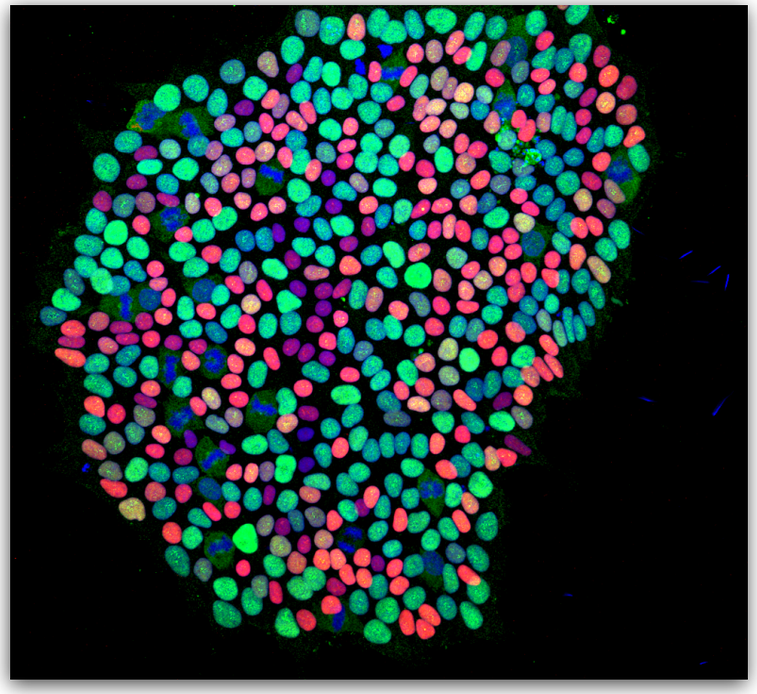
- $K =$ the number of components, all genes
- $\{f_{g,k}\} =$ gene-specific components
- mixing proportions in first condition, all genes

$$\phi = (\phi_1, \phi_2, \cdots, \phi_K)$$

- mixing proportions in second condition, all genes

$$\psi = (\psi_1, \psi_2, \cdots, \psi_K)$$

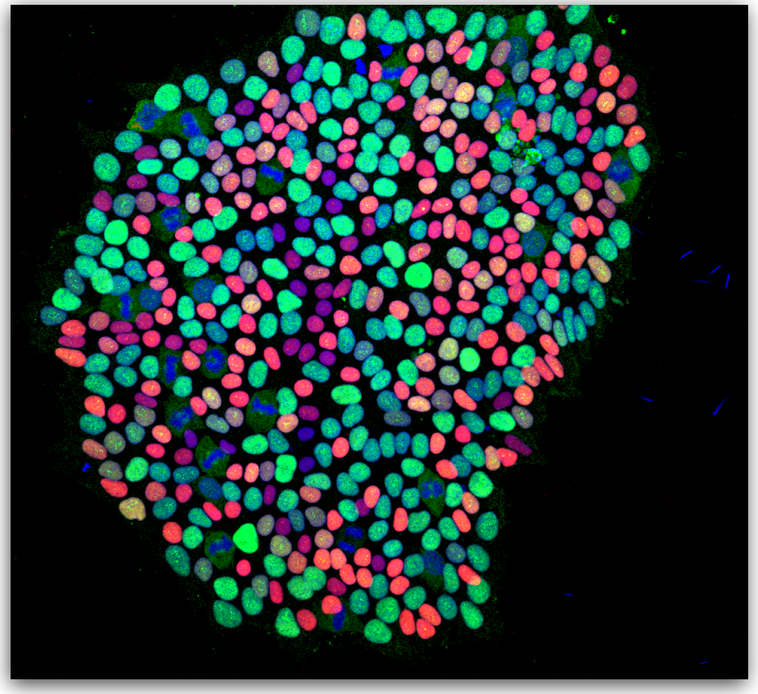
Compositional model



I.e.,

- ∃ K cell types (latent subpopulations)
- cell-level mixing induces mixing per gene
- conditions may use the subpopulations in different frequencies
- differences between conditions are due entirely to changes in subtype frequencies

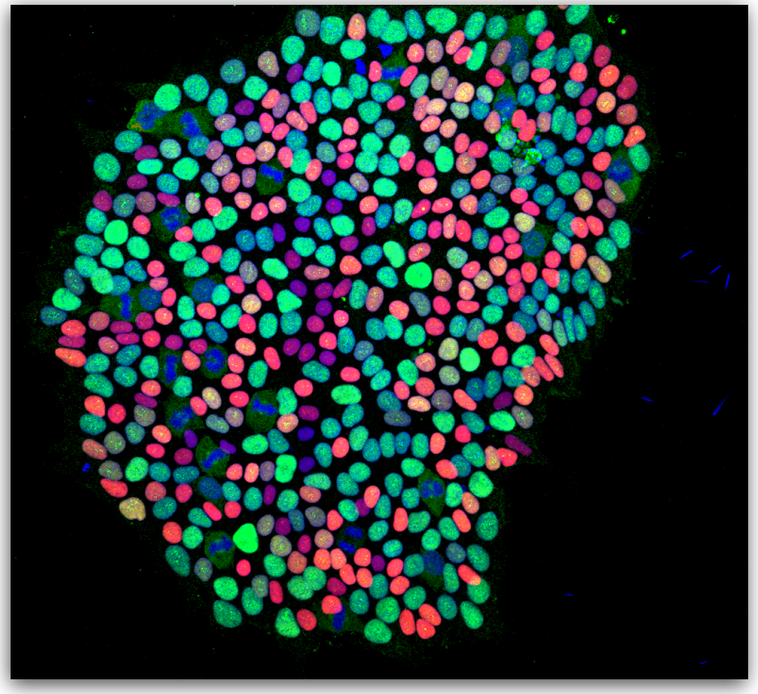
Compositional model



$$f_g^1(x) = \sum_{k=1}^{K} \phi_k f_{g,k}(x)$$

$$f_g^2(x) = \sum_{k=1}^{K} \psi_k f_{g,k}(x)$$

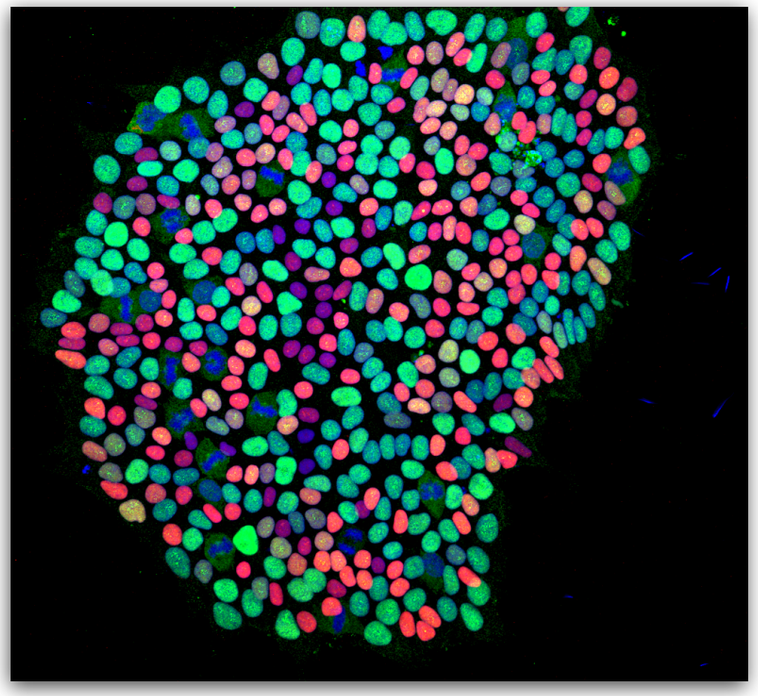
Compositional model



$$f_g^1(x) = \sum_{k=1}^{K} \phi_k f_{g,k}(x)$$

$$f_g^2(x) = \sum_{k=1}^{K} \psi_k f_{g,k}(x)$$

Compositional model

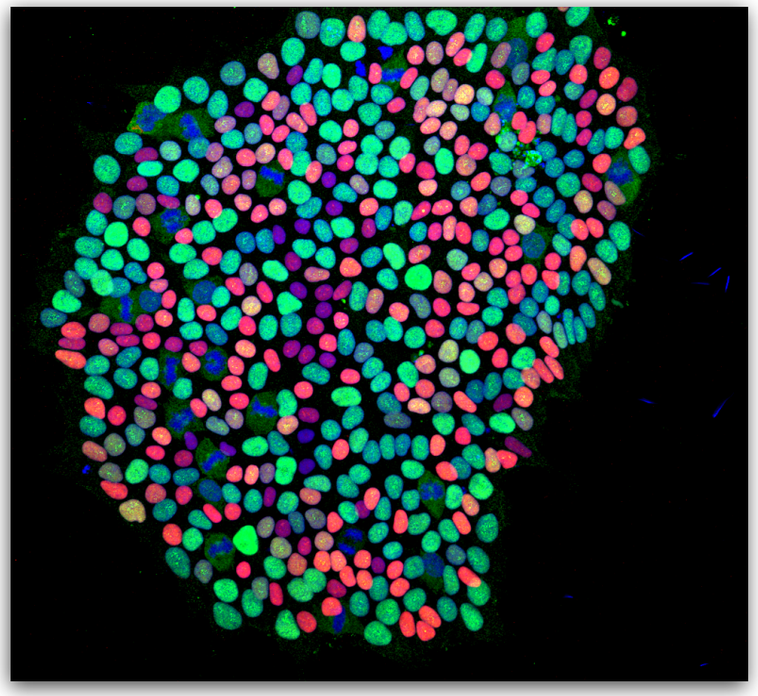


$$f_g^1(x) = \sum_{k=1}^{K} \phi_k f_{g,k}(x)$$

$$f_g^2(x) = \sum_{k=1}^{K} \psi_k f_{g,k}(x)$$

common to all genes

Compositional model



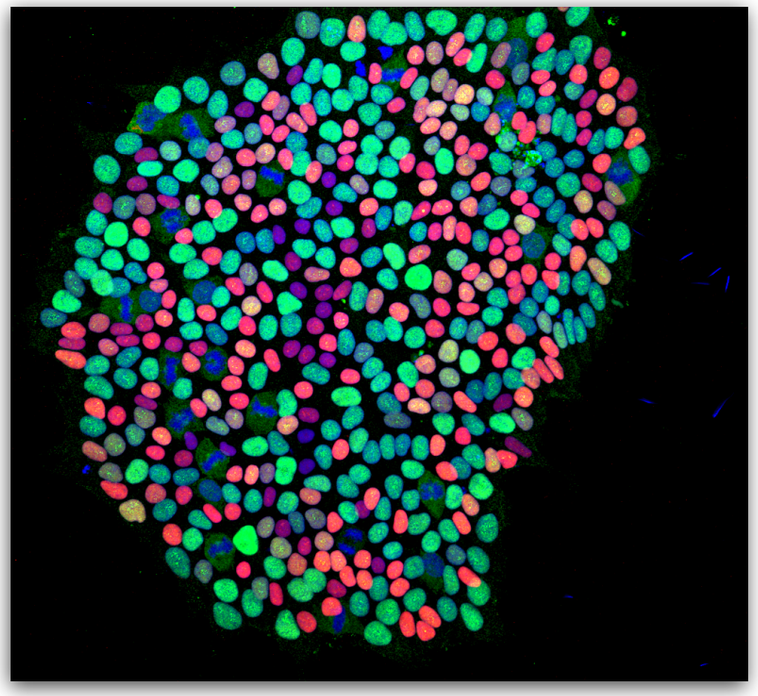
$$f_g^1(x) = \sum_{k=1}^{K} \phi_k f_{g,k}(x)$$

$$f_g^2(x) = \sum_{k=1}^{K} \psi_k f_{g,k}(x)$$

common to all genes

informed by whole genome data

Compositional model

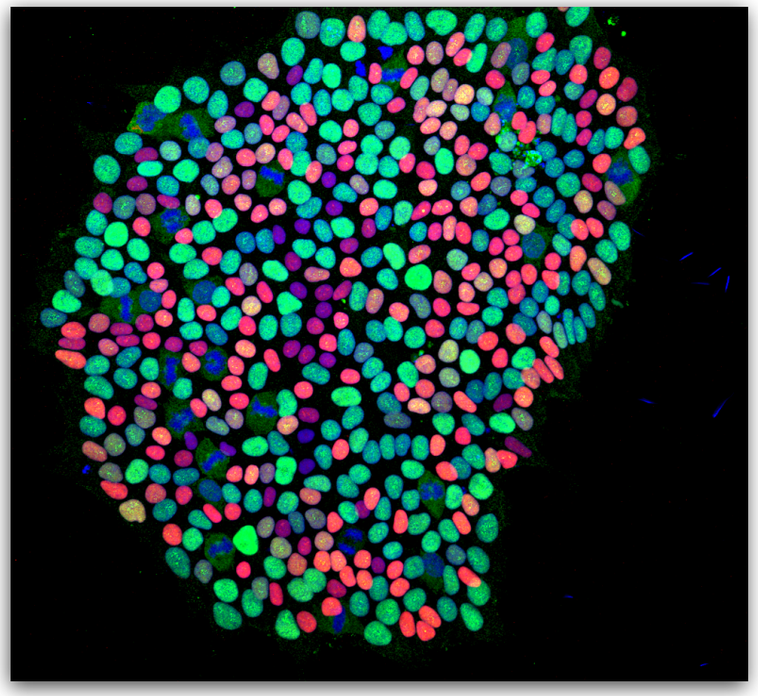


$$f_g^1(x) = \sum_{k=1}^{K} \phi_k f_{g,k}(x)$$

$$f_g^2(x) = \sum_{k=1}^{K} \psi_k f_{g,k}(x)$$

common to both conditions

Compositional model



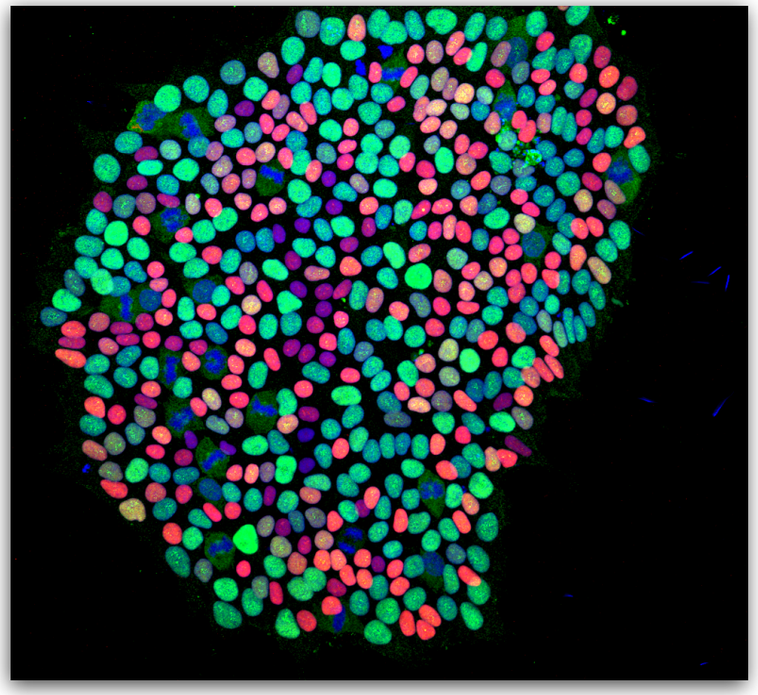
$$f_g^1(x) = \sum_{k=1}^{K} \phi_k f_{g,k}(x)$$

$$f_g^2(x) = \sum_{k=1}^{K} \psi_k f_{g,k}(x)$$

common to both conditions

informed by gene-level data

Compositional model



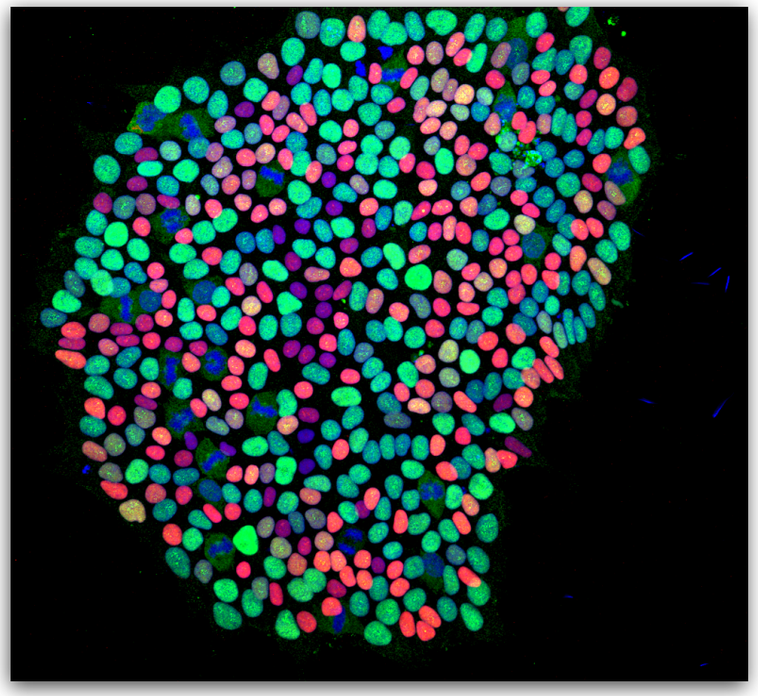
$$f_g^1(x) = \sum_{k=1}^{K} \phi_k f_{g,k}(x)$$

$$f_g^2(x) = \sum_{k=1}^{K} \psi_k f_{g,k}(x)$$

The parameters are: $\theta = (\phi, \psi, \text{parameters in } f_{g,k})$

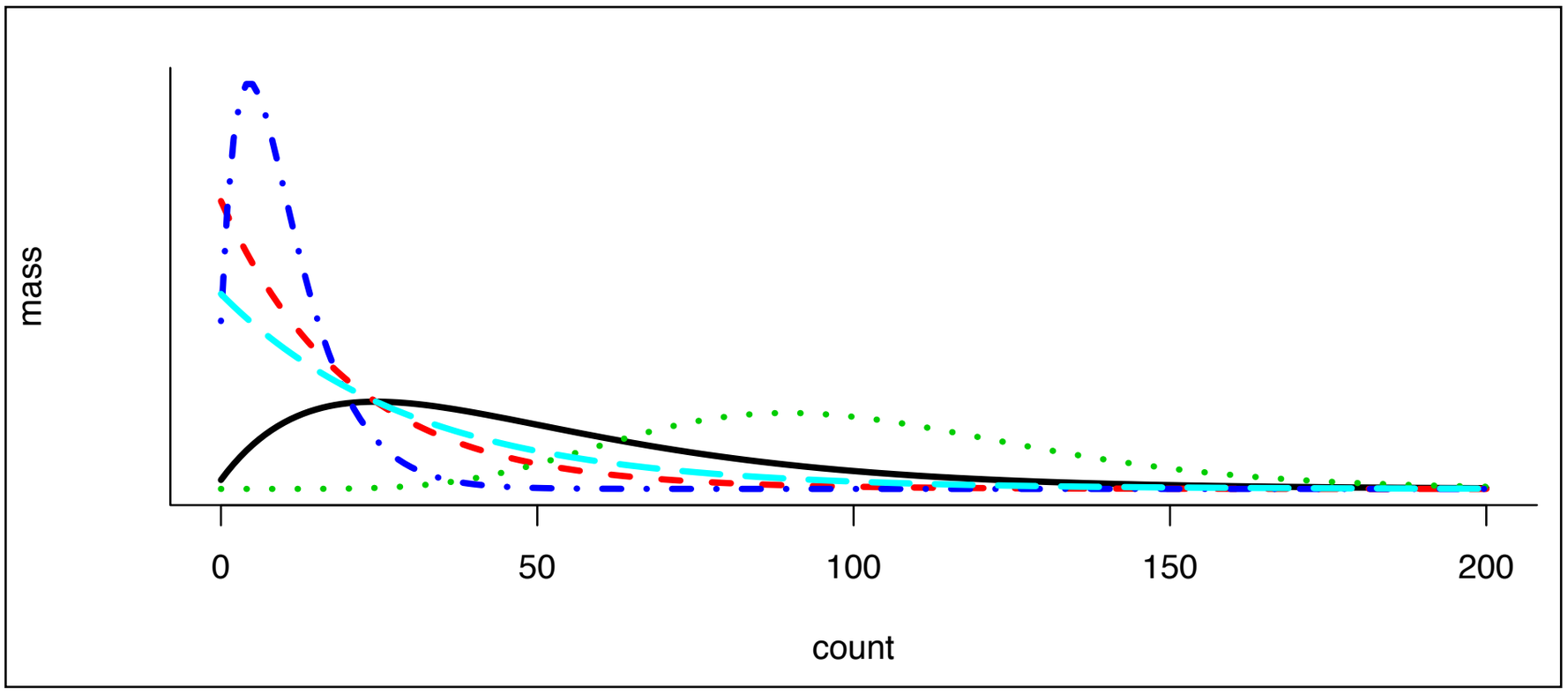
Component distributions

$$\{f_{g,k}\}$$



In all experiments so far,

$$f_{g,k} = \text{Negative Binomial}\left(\text{mean} = \mu_{g,k}, \text{shape} = \sigma_g\right)$$



often used for sequence counts:

EBSeq

edgeR

DESeq2

Gene-level inference

null

$$\text{Equivalent Distribution}_g = \text{ED}_g \qquad f_g^1(x) = f_g^2(x) \quad \forall x$$

alternative

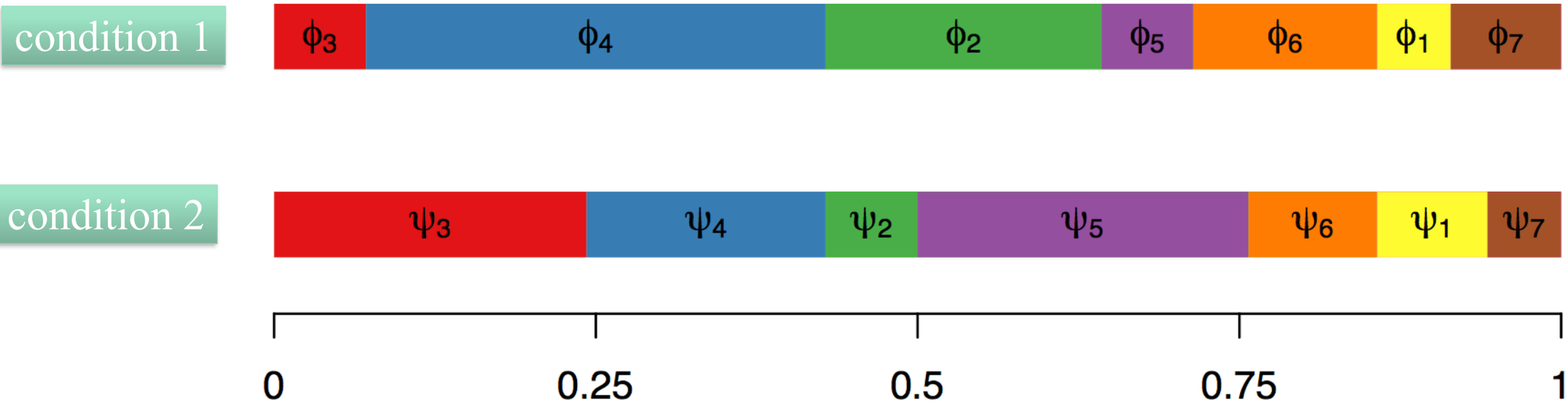
$$\text{Differential Distribution}_g = \text{DD}_g \qquad f_g^1(x) \neq f_g^2(x) \quad \text{for some } x$$

Important technical issue

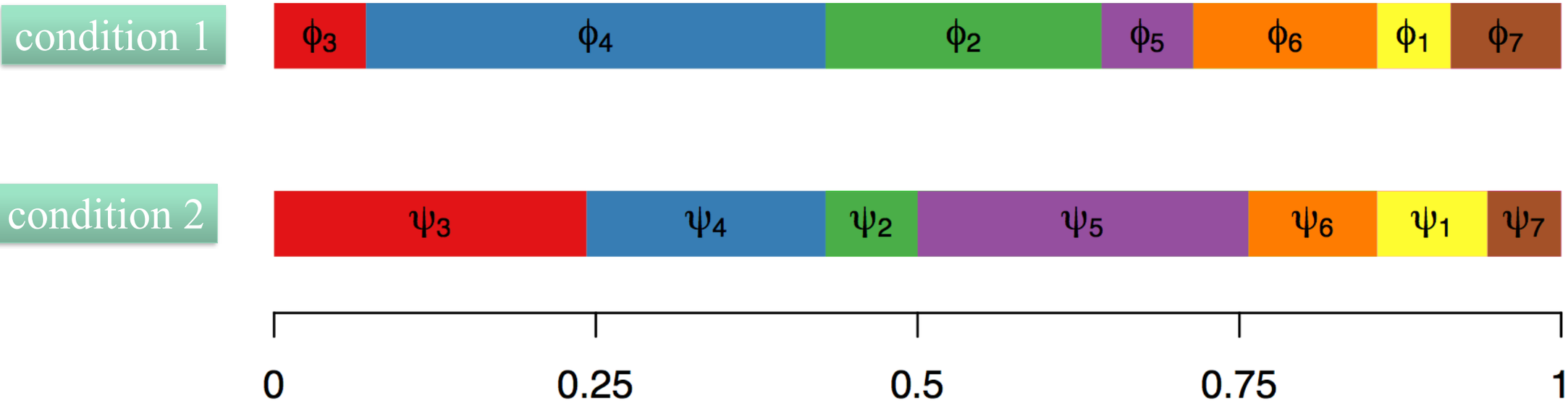
Differences in sub-type proportions may not change margins:

$$\phi \neq \psi \quad \not\Longrightarrow \quad \mathrm{DD}_g$$

Eg: frequencies of K=7 sub-types over two conditions



Eg: frequencies of K=7 sub-types over two conditions



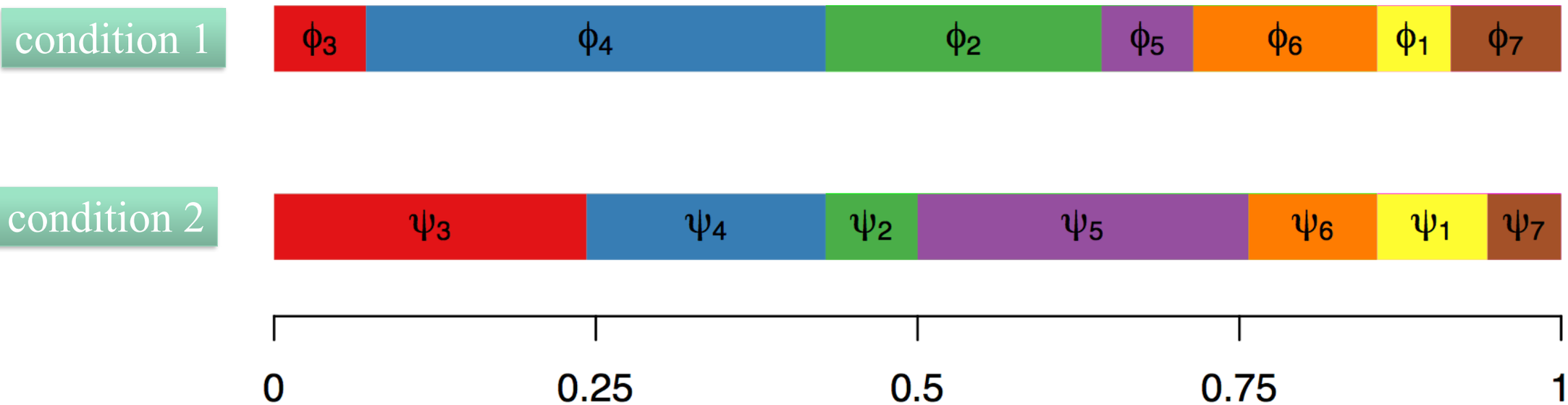
Eg: suppose, at this gene, some components are the same

$$f_{g,3} = f_{g,4} = \alpha$$

$$f_{g,2} = f_{g,5} = f_{g,6} = \beta$$

$$f_{g,1} = f_{g,7} = \gamma$$

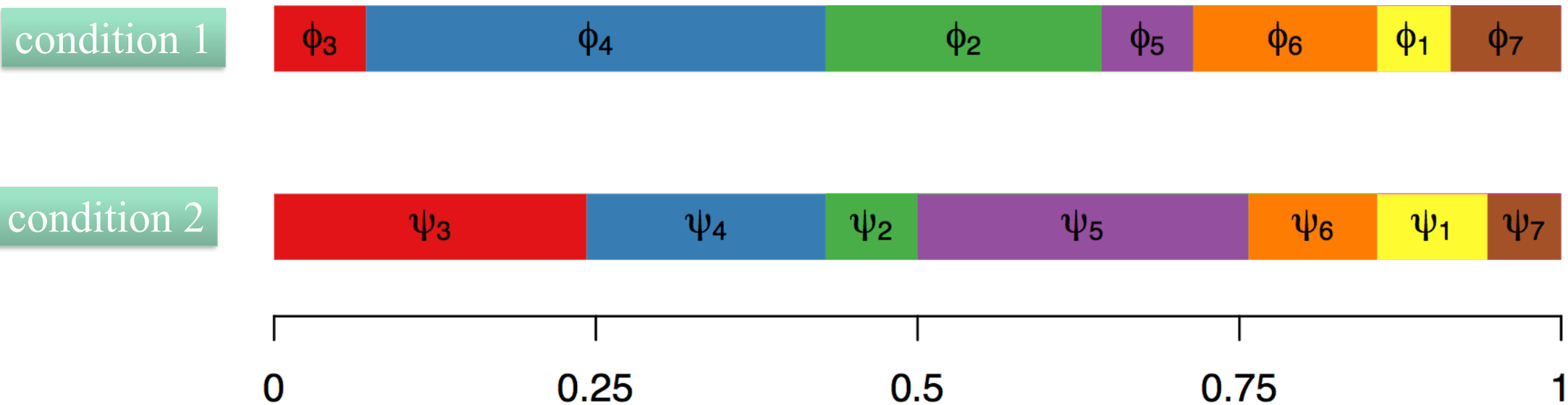
Eg: frequencies of K=7 sub-types over two conditions



Margins:

$$f_g^1(x) = \sum_{k=1}^{7} \phi_k f_{g,k}(x)$$

Eg: frequencies of K=7 sub-types over two conditions

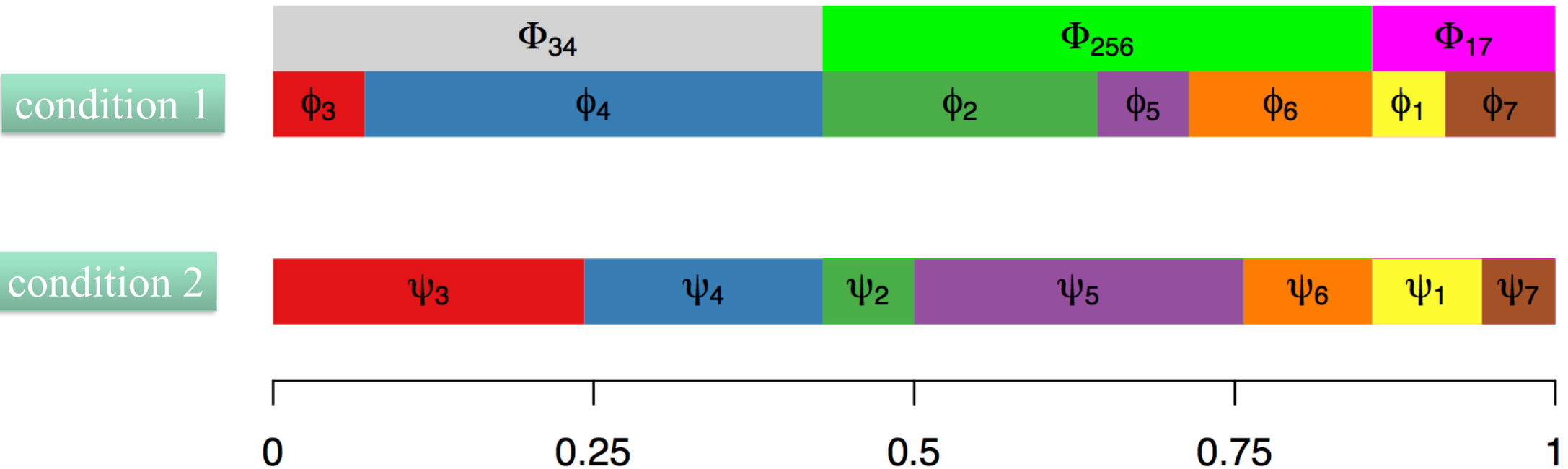


Margins:

$$f_g^1(x) = \sum_{k=1}^{7} \phi_k f_{g,k}(x)$$

$$= (\phi_3 + \phi_4)\alpha(x) + (\phi_2 + \phi_5 + \phi_6)\beta(x) + (\phi_1 + \phi_7)\gamma(x)$$

Eg: frequencies of K=7 sub-types over two conditions



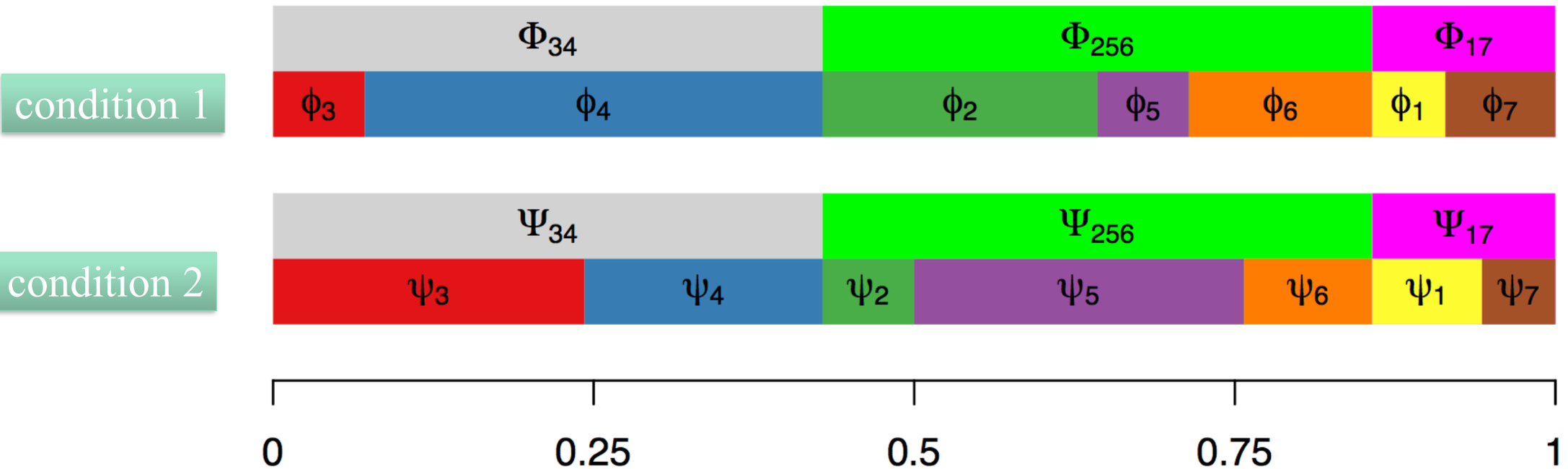
Margins:

$$f_g^1(x) = \sum_{k=1}^{7} \phi_k f_{g,k}(x)$$

$$= (\phi_3 + \phi_4)\alpha(x) + (\phi_2 + \phi_5 + \phi_6)\beta(x) + (\phi_1 + \phi_7)\gamma(x)$$

$$= \Phi_{3,4}\,\alpha(x) + \Phi_{2,5,6}\,\beta(x) + \Phi_{1,7}\,\gamma(x)$$

Eg: frequencies of K=7 sub-types over two conditions



Margins:

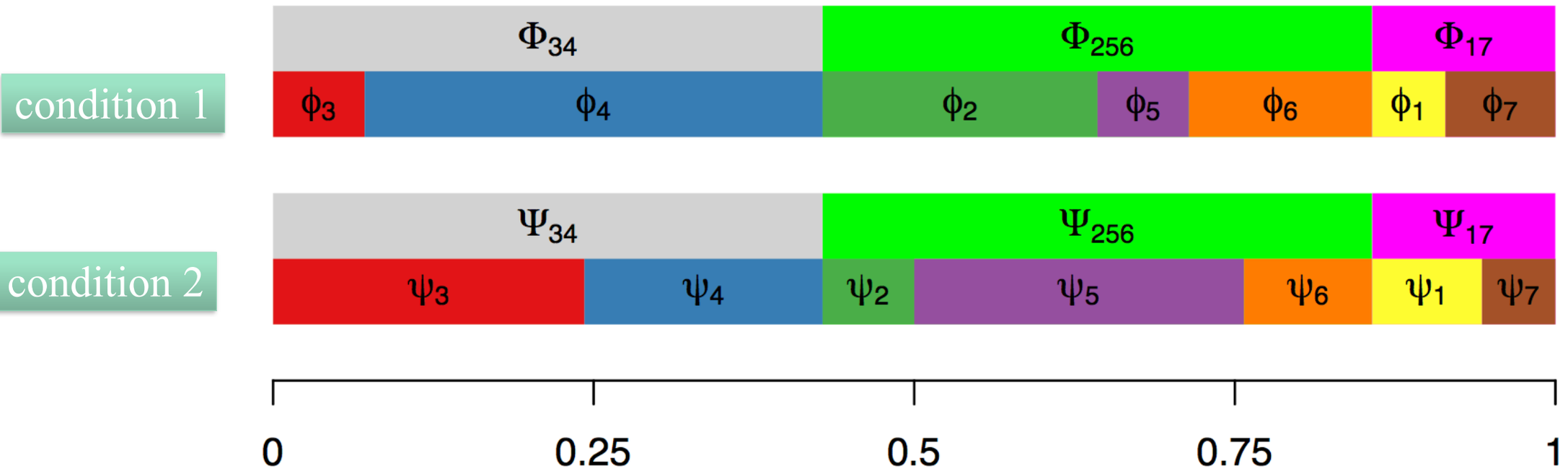
$$f_g^1(x) = \sum_{k=1}^{7} \phi_k f_{g,k}(x)$$

$$= (\phi_3 + \phi_4)\alpha(x) + (\phi_2 + \phi_5 + \phi_6)\beta(x) + (\phi_1 + \phi_7)\gamma(x)$$

$$= \Phi_{3,4}\,\alpha(x) + \Phi_{2,5,6}\,\beta(x) + \Phi_{1,7}\,\gamma(x)$$

$$= \Psi_{3,4}\,\alpha(x) + \Psi_{2,5,6}\,\beta(x) + \Psi_{1,7}\,\gamma(x)$$

Eg: frequencies of K=7 sub-types over two conditions



Margins:

$$
\begin{aligned}
f_g^1(x) &= \sum_{k=1}^{7} \phi_k f_{g,k}(x) \\
&= (\phi_3 + \phi_4)\alpha(x) + (\phi_2 + \phi_5 + \phi_6)\beta(x) + (\phi_1 + \phi_7)\gamma(x) \\
&= \Phi_{3,4}\,\alpha(x) + \Phi_{2,5,6}\,\beta(x) + \Phi_{1,7}\,\gamma(x) \\
&= \Psi_{3,4}\,\alpha(x) + \Psi_{2,5,6}\,\beta(x) + \Psi_{1,7}\,\gamma(x) \quad = f_g^2(x)
\end{aligned}
$$

Key fact: If *g* is has the same component distributions among two (or more) sub-types, and if these sub-types change in relative frequency between conditions such that their combined relative frequency does not change, then *g* has *equal distributions* (ED) between conditions

Key construct: sub-type partition

$$\pi = \{b\}$$

$$b \subset \{1, 2, \cdots, K\}$$

Block frequencies:

$$\Phi_b = \sum_{k \in b} \phi_k \qquad \Psi_b = \sum_{k \in b} \psi_k$$

Key construct: sub-type partition

$$\pi = \{b\}$$

$$b \subset \{1, 2, \cdots, K\}$$

Block frequencies:

$$\Phi_b = \sum_{k \in b} \phi_k \qquad \Psi_b = \sum_{k \in b} \psi_k$$

E.g. $\pi = \{\{3, 4\}, \{2, 5, 6\}, \{1, 7\}\}$

Key parameter subsets:

$$A_\pi = \{\theta \in \Theta : \ \Phi_b = \Psi_b \, \forall b \in \pi\}$$

$$M_{g,\pi} = \{\theta \in \Theta : \ \mu_{g,k} = \mu_{g,k'} \iff k, k' \in b, b \in \pi\}$$

E.g. $\pi = \{\{3,4\}, \{2,5,6\}, \{1,7\}\}$

Key parameter subsets:

$$A_\pi = \{\theta \in \Theta : \ \Phi_b = \Psi_b \, \forall b \in \pi\}$$

subtype frequencies, both conditions

$$M_{g,\pi} = \{\theta \in \Theta : \ \mu_{g,k} = \mu_{g,k'} \iff k, k' \in b, b \in \pi\}$$

expression changes between subtypes

E.g. $\pi = \{\{3,4\}, \{2,5,6\}, \{1,7\}\}$

Theorem: $\mathrm{ED}_g = \bigcup_{\pi \in \Pi} (A_\pi \cap M_{g,\pi})$

all partitions of K subtypes

Theorem: $$\mathrm{ED}_g = \bigcup_{\pi \in \Pi} (A_\pi \cap M_{g,\pi})$$

all partitions of K subtypes

local FDR

$$1 - P(\mathrm{DD}_g|X, y) = P(\mathrm{ED}_g|X, y) = \sum_{\pi \in \Pi} P\left(A_\pi \cap M_{g,\pi}|X, y\right)$$

Theorem: $$\mathrm{ED}_g = \bigcup_{\pi \in \Pi} (A_\pi \cap M_{g,\pi})$$

all partitions of K subtypes

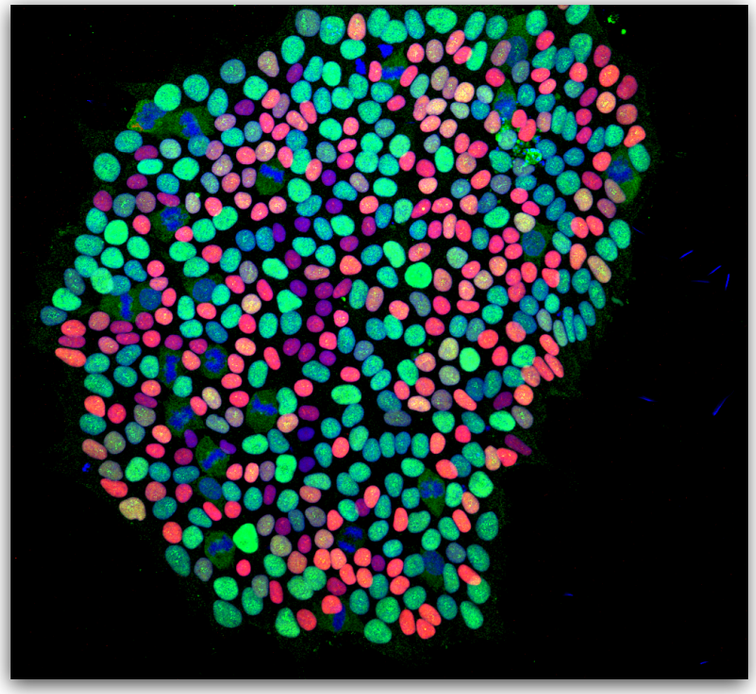
local FDR

$$1 - P(\mathrm{DD}_g | X, y) = P(\mathrm{ED}_g | X, y) = \sum_{\pi \in \Pi} P\left(A_\pi \cap M_{g,\pi} | X, y\right)$$

Latent variables

$$z_c = \text{subtype label of cell } c$$

$$z = \{z_c\}$$

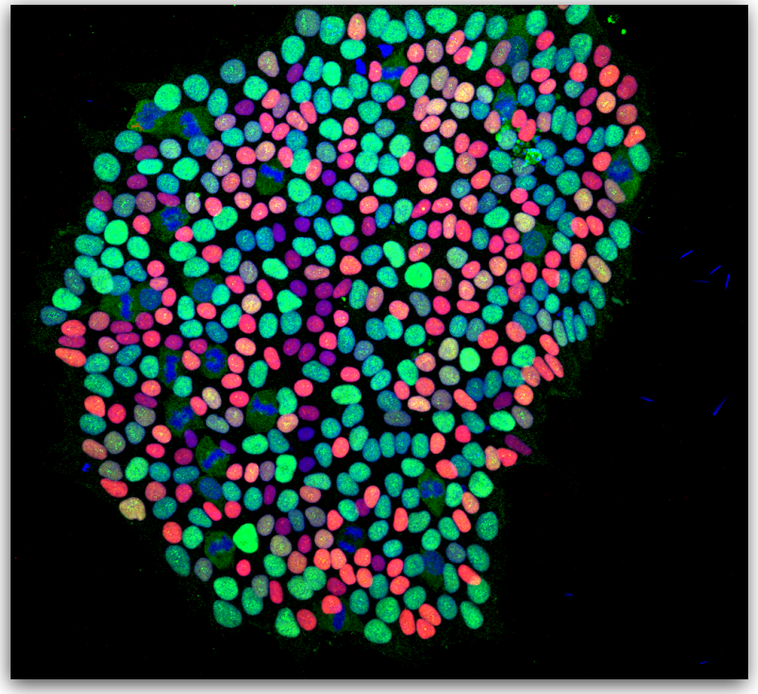


*under prior independence

Latent variables

$z_c = \text{subtype label of cell } c$

$z = \{z_c\}$



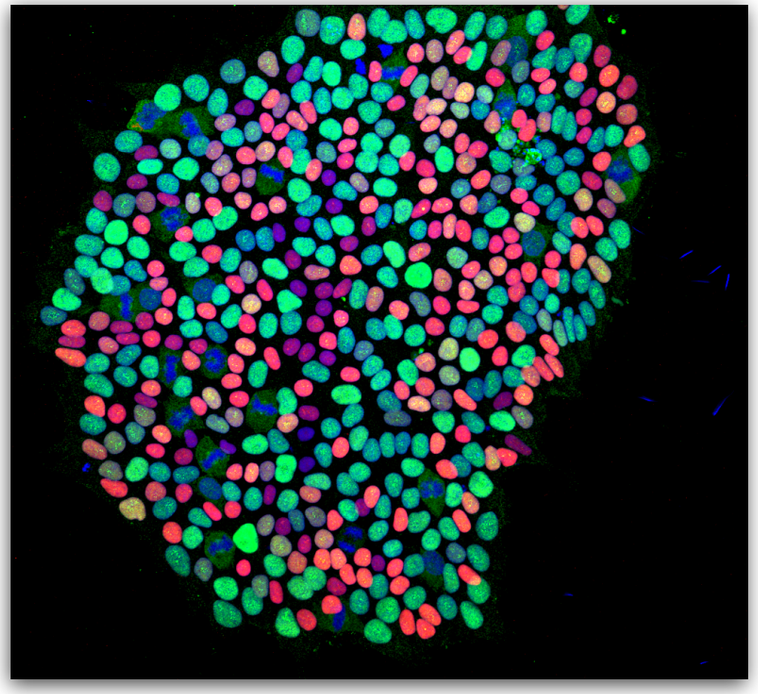
*under prior independence

Theorem: $P\left(A_\pi \cap M_{g,\pi} | X, y, z\right) = P\left(A_\pi | y, z\right) \; P\left(M_{g,\pi} | X, z\right)$

Latent variables

$$z_c = \text{subtype label of cell } c$$

$$z = \{z_c\}$$



*under prior independence

Theorem: $$P\left(A_\pi \cap M_{g,\pi} | X, y, z\right) = P\left(A_\pi | y, z\right) \; P\left(M_{g,\pi} | X, z\right)$$

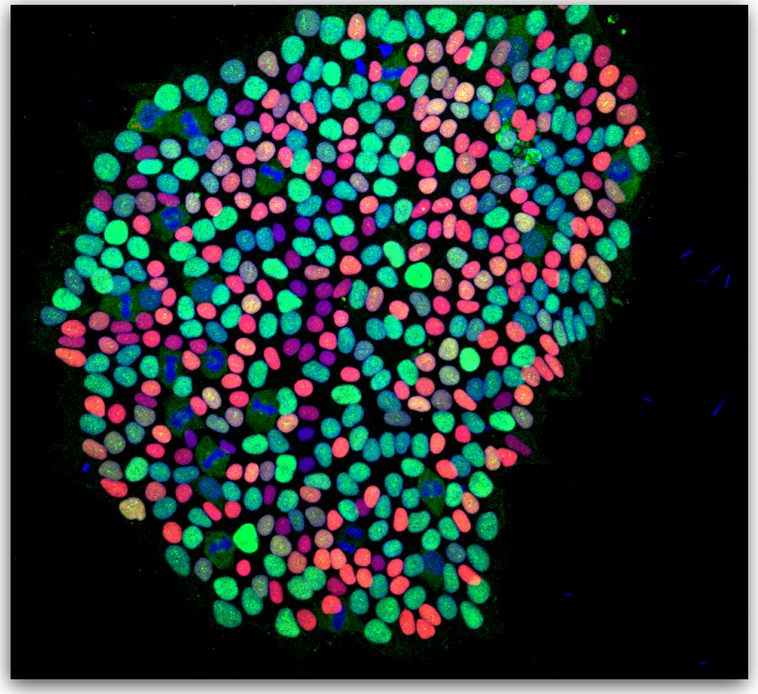
Idea 1: use clustering to estimate: $\hat{z}$

$$P\left(A_\pi \cap M_{g,\pi} | X, y\right) \approx P\left(A_\pi \cap M_{g,\pi} | X, y, \hat{z}\right)$$

Latent variables

$z_c = \text{subtype label of cell } c$

$z = \{z_c\}$



*under prior independence

Theorem: $P\left(A_\pi \cap M_{g,\pi} | X, y, z\right) = P\left(A_\pi | y, z\right) \; P\left(M_{g,\pi} | X, z\right)$

Idea 1: use clustering to estimate: $\hat{z}$

$$P\left(A_\pi \cap M_{g,\pi} | X, y\right) \approx P\left(A_\pi \cap M_{g,\pi} | X, y, \hat{z}\right)$$

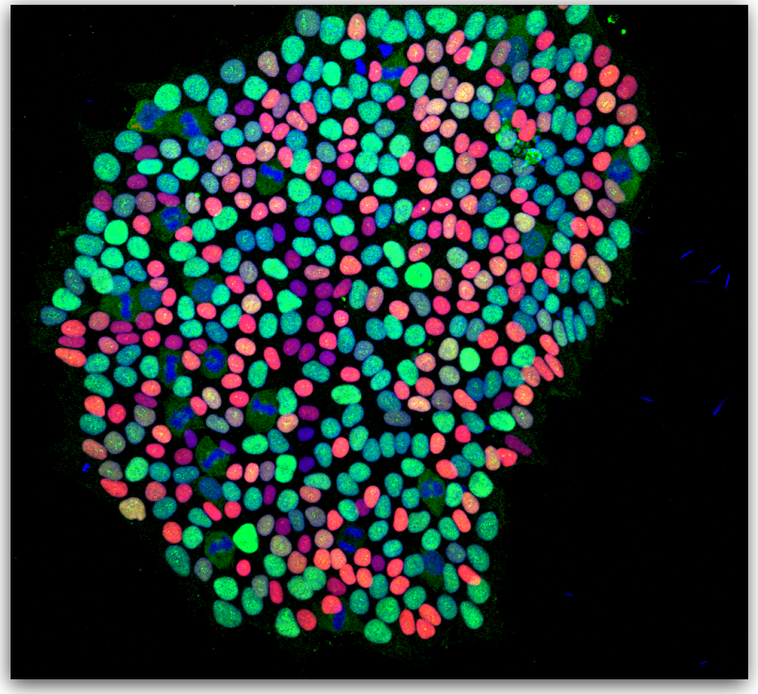
e.g.

SC3: consensus clustering of single-cell RNA-seq data *Nature Methods* **14**, 483–486 (2017)

Latent variables

$z_c = \text{subtype label of cell } c$

$z = \{z_c\}$



*under prior independence

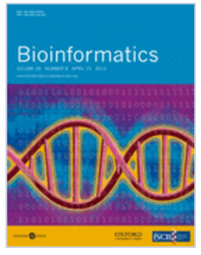
Theorem: $P\left(A_\pi \cap M_{g,\pi} | X, y, z\right) = P\left(A_\pi | y, z\right) \; P\left(M_{g,\pi} | X, z\right)$

Idea 2: average conditional probabilities over z's obtained by clustering with randomized distances

smoothing/stability

$$P\left(M_{g,\pi}|X,z\right)$$

expression changes between subtypes



Volume 29, Issue 8
15 April 2013

EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments

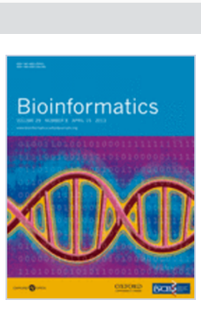
Ning Leng; John A. Dawson; James A. Thomson; Victor Ruotti; Anna I. Rissman; Bart M. G. Smits; Jill D. Haag; Michael N. Gould; Ron M. Stewart; Christina Kendziorski

$$P\left(M_{g,\pi}|X,z\right)$$

expression changes between subtypes

$$P(A_{\pi}|y,z)$$

subtype frequency changes between conditions



Volume 29, Issue 8
15 April 2013

EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments

Ning Leng; John A. Dawson; James A. Thomson; Victor Ruotti; Anna I. Rissman; Bart M. G. Smits; Jill D. Haag; Michael N. Gould; Ron M. Stewart; Christina Kendziorski

Double-Dirichlet prior

Double Dirichlet

$$p(\phi, \psi) = \sum_{\pi \in \Pi} P(A_\pi)\, p(\phi, \psi | A_\pi)$$

$$\Phi_\pi \sim \text{Dirichet}_{N(\pi)}[\beta_\pi] \qquad \Psi_\pi = \Phi_\pi$$

$$\tilde{\phi}_b = \frac{\phi_b}{\Phi_b} \qquad \tilde{\psi}_b = \frac{\psi_b}{\Psi_b}$$

$$\tilde{\phi}_b, \tilde{\psi}_b \sim_{\text{i.i.d.}} \text{Dirichlet}_{N(b)}[\alpha_b]$$

Double Dirichlet

$$p(A_\pi|y,z) \propto \sum_{\pi' \in RF(\pi)} w(\pi',\pi')$$

$$w(\pi,\pi) = p(t^1|t^1_\pi)\, p(t^2|t^2_\pi)\, p(t^1_\pi, t^2_\pi|A_\pi) p(A_\pi).$$

(2) $$p(t^j|t^j_\pi) = \prod_{b\in\pi} \left\{ \left[\frac{\Gamma(t^j_b+1)}{\prod_{k\in b}\Gamma(t^j_k+1)} \right] \left[\frac{\Gamma(\sum_{k\in b}\alpha_k)}{\prod_{k\in b}\Gamma(\alpha_k)} \right] \left[\frac{\prod_{k\in b}\Gamma(\alpha_k+t^j_k)}{\Gamma(t^j_b+\sum_{k\in b}\alpha_k))} \right] \right\}$$

and

(3) $$p(t^1_\pi, t^2_\pi|A_\pi) = \left[\frac{\Gamma(n_1+1)\Gamma(n_2+1)}{\prod_{b\in\pi}\Gamma(t^1_b+1)\Gamma(t^2_b+1)} \right] \left[\frac{\Gamma(\sum_{b\in\pi}\beta_b)}{\prod_{b\in\pi}\Gamma(\beta_b)} \right] \left[\frac{\prod_{b\in\pi}\Gamma(\beta_b+t^1_b+t^2_b)}{\Gamma(n_1+n_2+\sum_{b\in\pi}\beta_b)} \right].$$

Double Dirichlet

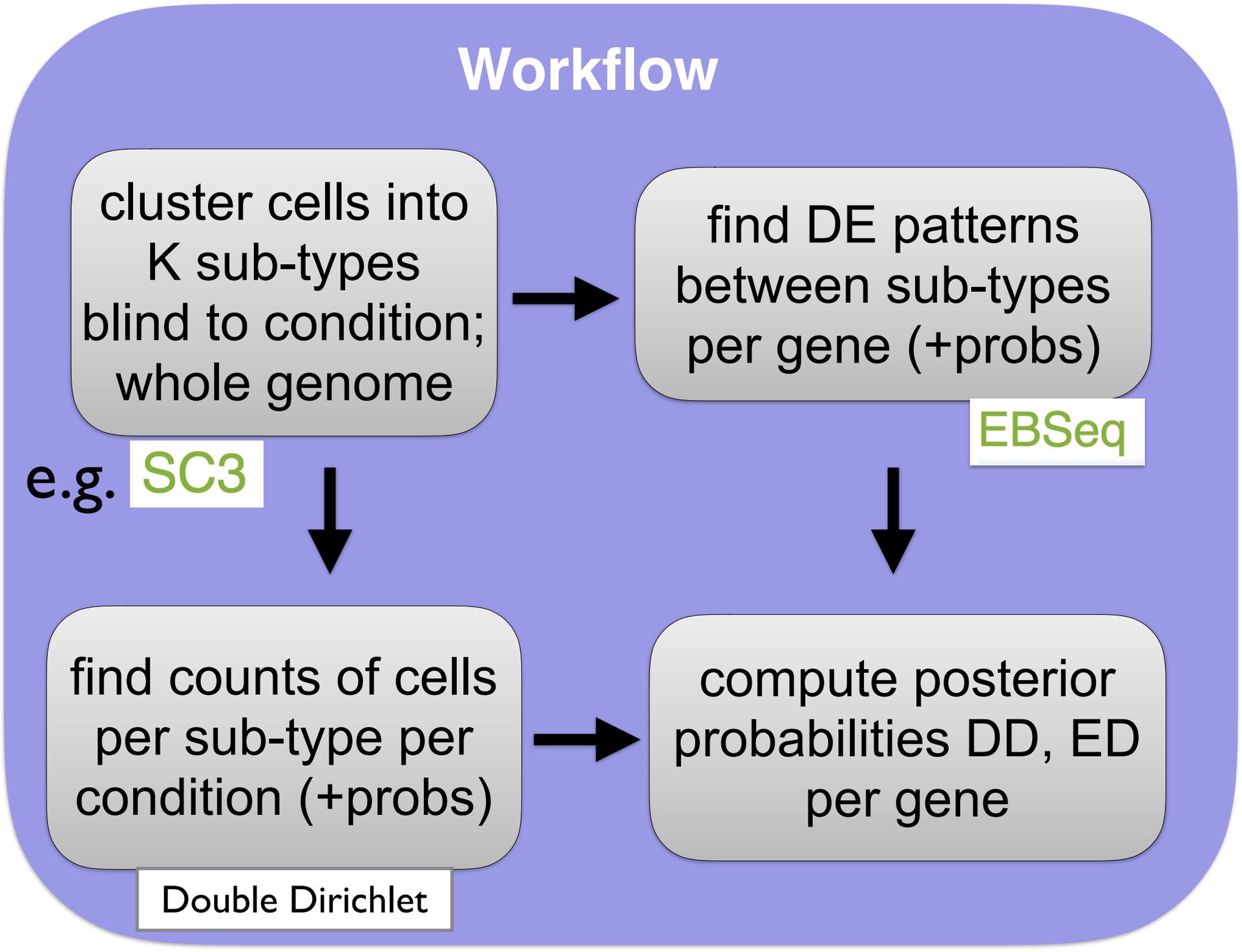
$$p(A_\pi|y,z) \propto \sum_{\pi' \in RF(\pi)} w(\pi',\pi')$$

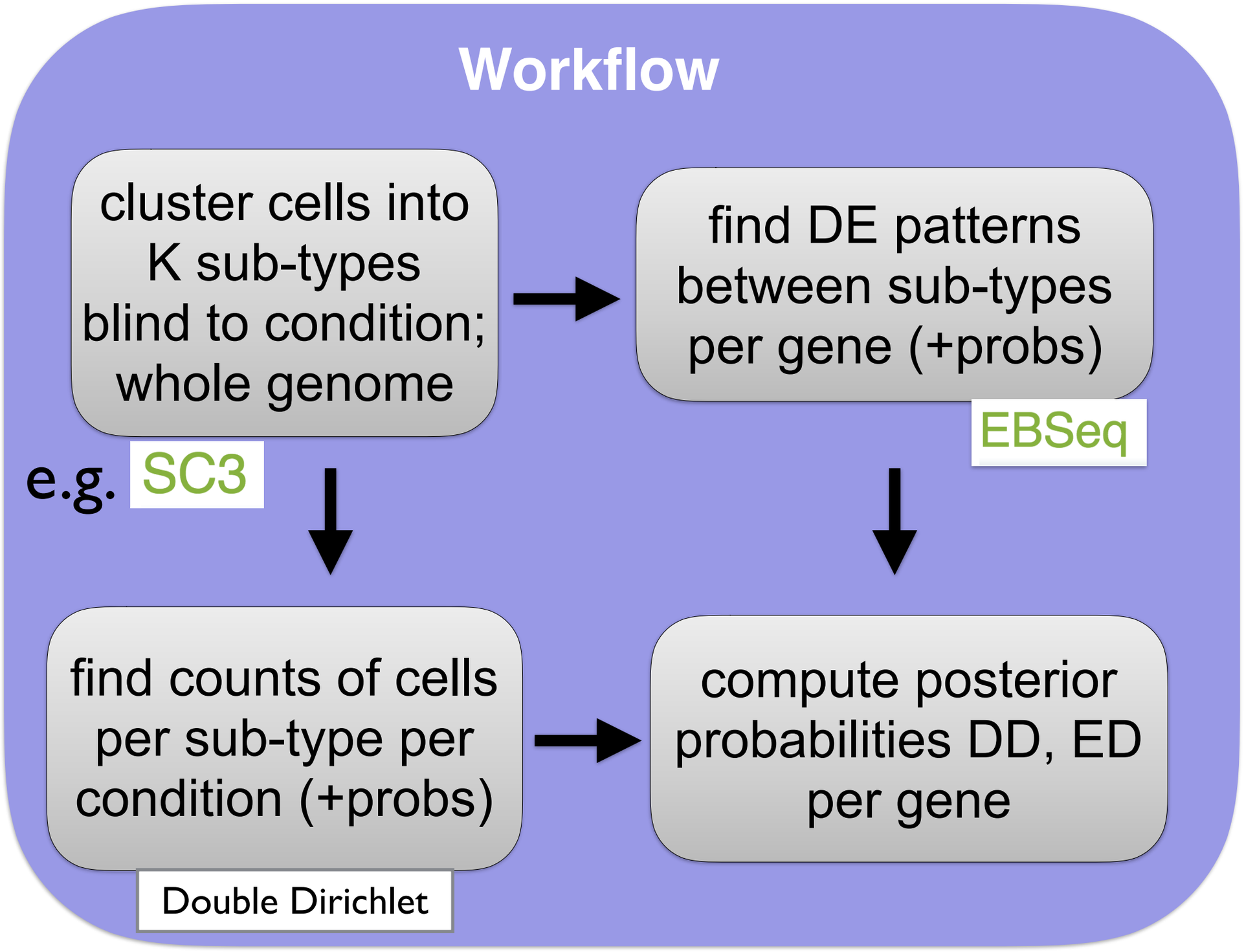
$$w(\pi,\pi) = p(t^1|t^1_\pi)\, p(t^2|t^2_\pi)\, p(t^1_\pi, t^2_\pi|A_\pi) p(A_\pi).$$

$$(2) \qquad p(t^j|t^j_\pi) = \prod_{b\in\pi} \left\{ \left[\frac{\Gamma(t^j_b+1)}{\prod_{k\in b}\Gamma(t^j_k+1)} \right] \left[\frac{\Gamma(\sum_{k\in b}\alpha_k)}{\prod_{k\in b}\Gamma(\alpha_k)} \right] \left[\frac{\prod_{k\in b}\Gamma(\alpha_k+t^j_k)}{\Gamma(t^j_b+\sum_{k\in b}\alpha_k))} \right] \right\}$$

and

$$(3) \quad p(t^1_\pi, t^2_\pi|A_\pi) = \left[\frac{\Gamma(n_1+1)\Gamma(n_2+1)}{\prod_{b\in\pi}\Gamma(t^1_b+1)\Gamma(t^2_b+1)} \right] \left[\frac{\Gamma(\sum_{b\in\pi}\beta_b)}{\prod_{b\in\pi}\Gamma(\beta_b)} \right] \left[\frac{\prod_{b\in\pi}\Gamma(\beta_b+t^1_b+t^2_b)}{\Gamma(n_1+n_2+\sum_{b\in\pi}\beta_b)} \right].$$


Workflow
cluster cells into K sub-types blind to condition; whole genome
e.g. SC3
find DE patterns between sub-types per gene (+probs)
EBSeq
find counts of cells per sub-type per condition (+probs)
Double Dirichlet
compute posterior probabilities DD, ED per gene



*repeat on randomized clusterings

Algorithm 1 scDDBoost-core

Input:

GENES by CELLS expression data matrix $X = (X_{g,c})$

cell condition labels $y = (y_c)$

cell subtype labels (estimated) $\hat{z}$

Output: posterior probabilities of differential distribution from estimated subtypes

1: **procedure** SCDDBOOST-CORE($X, y, \hat{z}$)

2: number of cell subtypes $K = \text{length}(\text{unique}(\hat{z}))$

3: subtype differential expression: $\forall g, \pi$ compute $P(M_{g,\pi}|X, \hat{z})$ using EBSeq[13]

4: cell frequency changes: $\forall \pi$ compute $P(A_\pi|y, \hat{z})$ using Double Dirichlet model

5: posterior probability: $\forall g,\ P(\text{ED}_g|X, y, \hat{z}) \leftarrow \sum_\pi P(M_{g,\pi}|X, \hat{z})\, P(A_\pi|y, \hat{z})$

6: **return** $\forall g,\ P(\text{DD}_g|X, y, \hat{z}) = 1 - P(\text{ED}_g|X, y, \hat{z})$

Algorithm 2 scDDboost

Input:

GENES by CELLS expression data matrix $X = (X_{g,c})$
cell condition labels $y = (y_c)$
number of cell subtypes K
number of randomized clusterings n_r
regularization parameter λ

Output: posterior probabilities of differential distribution

procedure SCDDBOOST(X, y, K, n_r, λ)
distance matrix: $D = \text{dist(X)} \leftarrow$ pairwise distances between cells (columns of X)
repeat
Exponential noise vector: e, with components $\sim \text{Exp}(\lambda)$
randomized distance matrix: $D^* \leftarrow D + e\mathbf{1}^T + \mathbf{1}e^T$
$P^* \leftarrow$ SCDDBOOST-CORE(X, y, D^*, K)
until n_r randomized distance matrices
return $\forall$genes g, $P(\text{DD}_g|X, y) = \frac{1}{n_r}\sum_{D^*} P_g^*$

Summary

- new method for scoring expression changes with scRNA-seq data

scDDboost

clustering/testing

Summary

- new method for scoring expression changes with scRNA-seq data **scDDboost**

  clustering/testing

- on estimating K

- on randomized clusterings

- behavior with large numbers of cells

- on applications to 10x data:
  tumor cell heterogeneity [Halberg]
  hematopoiesis [Bresnick]