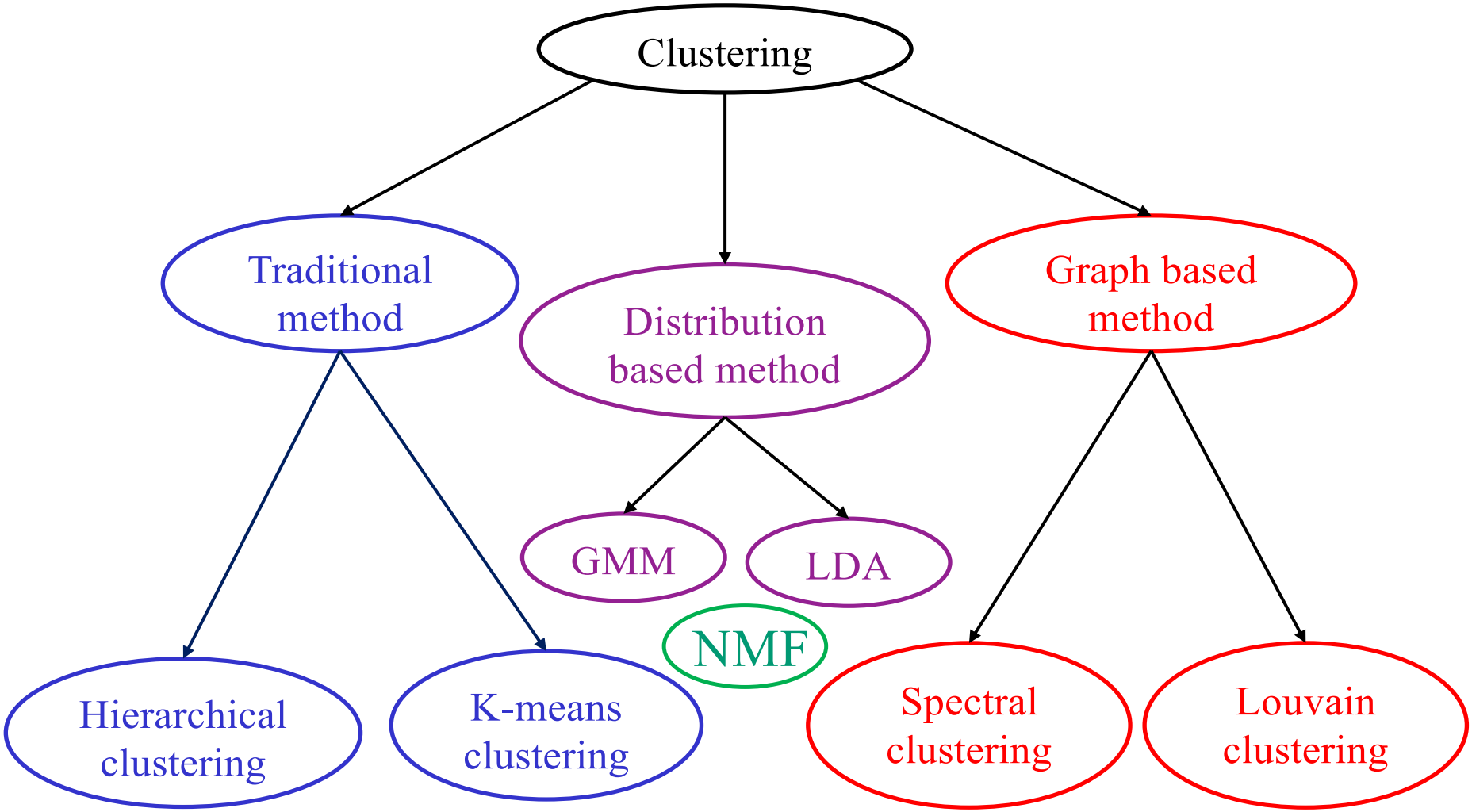
Clustering Methods:
From k-means to Gaussian Mixture Model and Louvain Algorithm

Maxwell Lee

High-dimension Data Analysis Group
Laboratory of Cancer Biology and Genetics
Center for Cancer Research
National Cancer Institute

January 25, 2021

Outline of Clustering Methods

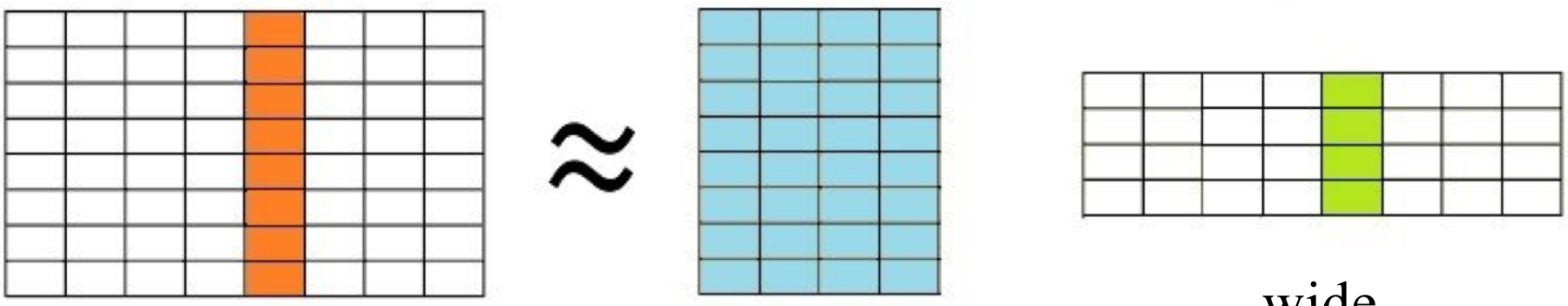


GMM: Gaussian Mixture Model
LDA: Latent Dirichlet Allocation
NMF: Non-negative matrix factorization

Contributed by Emily Tai

Mathematic Model of Non-Negative Matrix Factorization

X_{mn} W_{mk} H_{kn}



tall

wide

X_{mn}: m features; n samples
W_{mk}: m features; k latent variables
H_{km}: k latent variables; n encodings

latent variables:
basis images
topics
centroids
signatures

Each element of matrix is non-negative

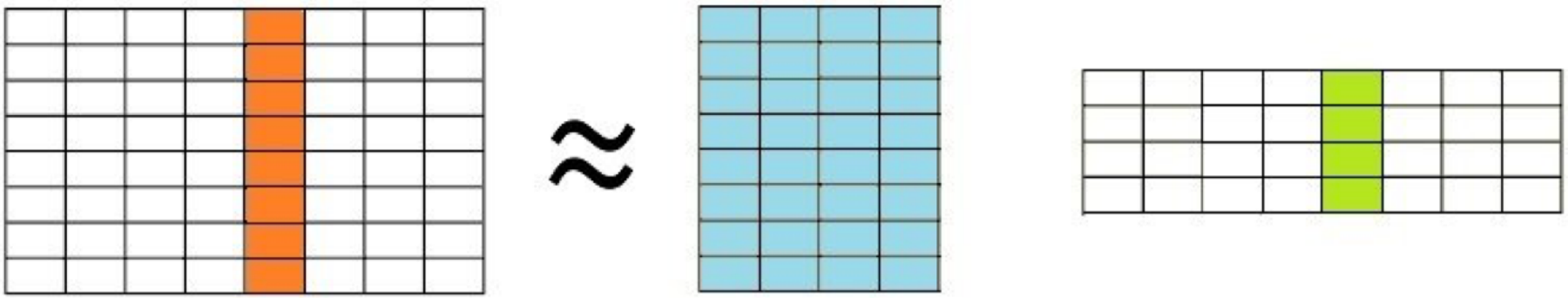
$X >= 0; W >= 0; H >= 0$

$k << \min(m,n)$ ⟶ Dimension reduction

Lee and Seung, Nature 1999; 401:788–791

Understanding NMF from Topic Modeling (Mixture Model)

X_{mn} W_{mk} H_{kn}



X_{mn}: m words; n documents
W_{mk}: m words; k topics
H_{km}: k topics; n documents

Topic modeling:
$\Sigma_i X_{ij} = 1$ (column sums to 1)
$\Sigma_i W_{ij} = 1$ and $\Sigma_i H_{ij} = 1$

$x_{ij} \sim \Sigma_k w_{ik} h_{kj}$

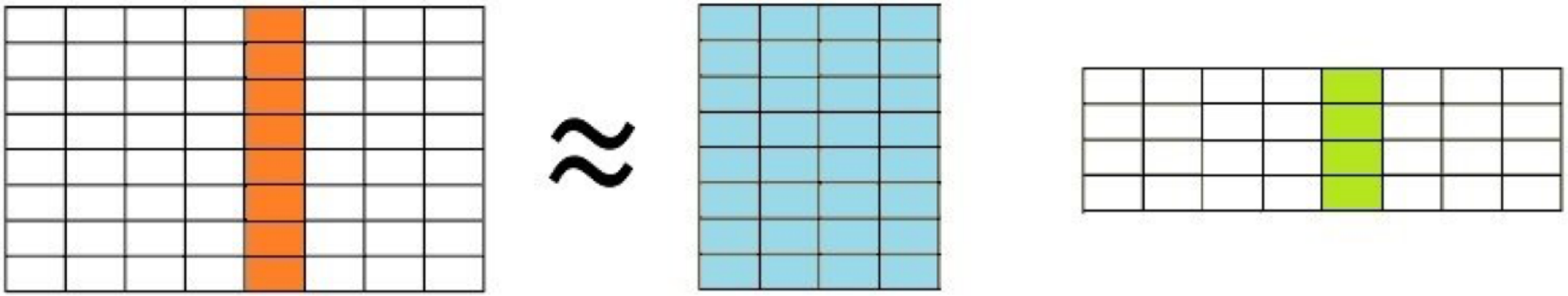
Understanding NMF from Topic Modeling (Mixture Model)

I will talk about spectral clustering, which is a graph-based method and consists of dimension reduction with Laplacian Eigenmap and k-means clustering in the reduced dimension space. I will also talk about Louvain algorithm, which is used in Seurat package to cluster single cell RNAseq data. Louvain algorithm is a network community approach. It is very fast and has capacity to do clustering analysis for million nodes in a network. I will provide practical examples to illustrate how each method works and how to interpret the results of clustering analysis and explain the pros and cons of each method.

H=0.4*clustering+0.2*dimension+0.2*algorithm+0.2*network

Understanding NMF from Topic Modeling (Mixture Model)

X_{mn} W_{mk} H_{kn}



Topic modeling:

$\Sigma_i X_{ij} = 1$ (column sums to 1)

$\Sigma_i W_{ij} = 1$ and $\Sigma_i H_{ij} = 1$

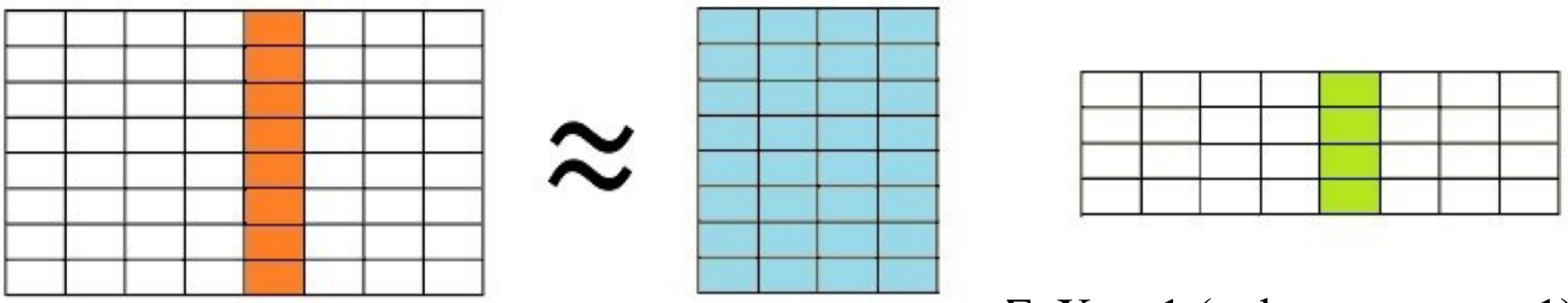
Limitations:

$n >> m$

Documents need to be large

Latent Dirichlet Allocation (LDA)

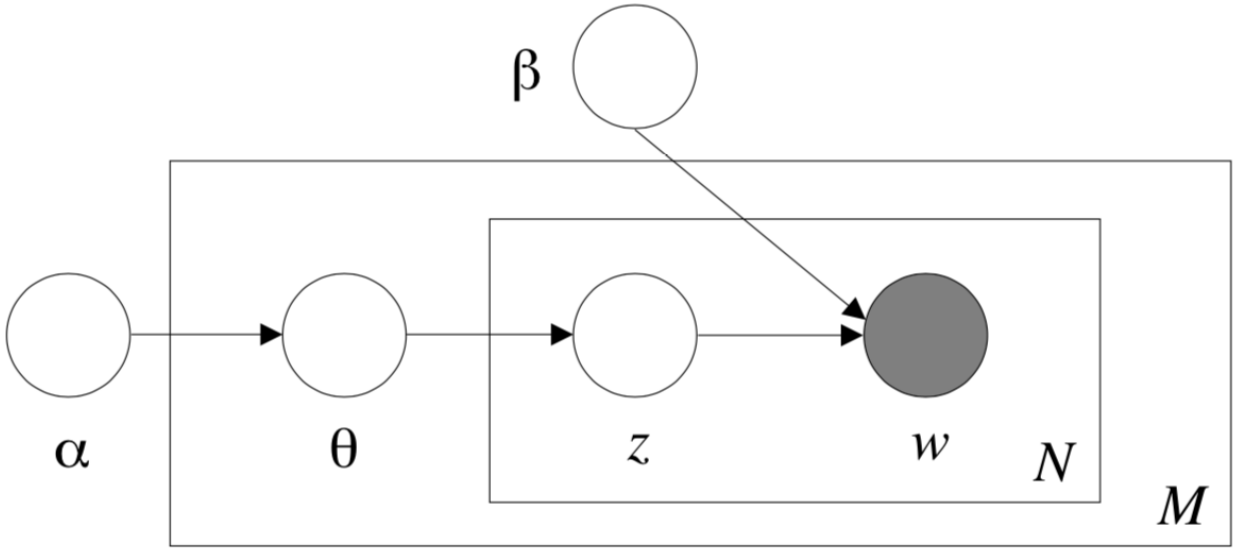
X_{mn} W_{mk} H_{kn}



$\Sigma_i X_{ij} = 1$ (column sums to 1)

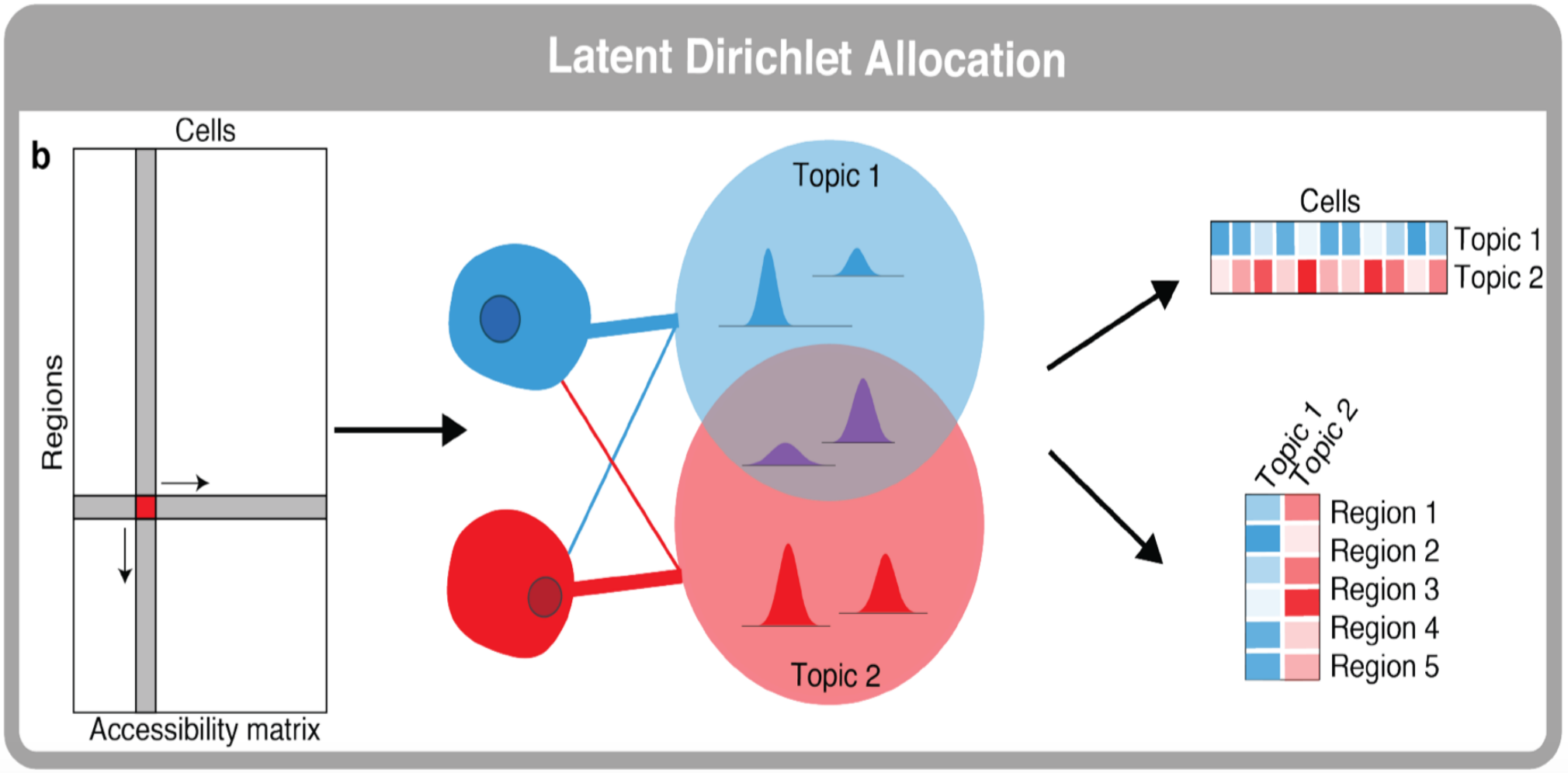
$\Sigma_i W_{ij} = 1$ and $\Sigma_i H_{ij} = 1$

Bayesian Hierarchical Model and Generative Statistical Model



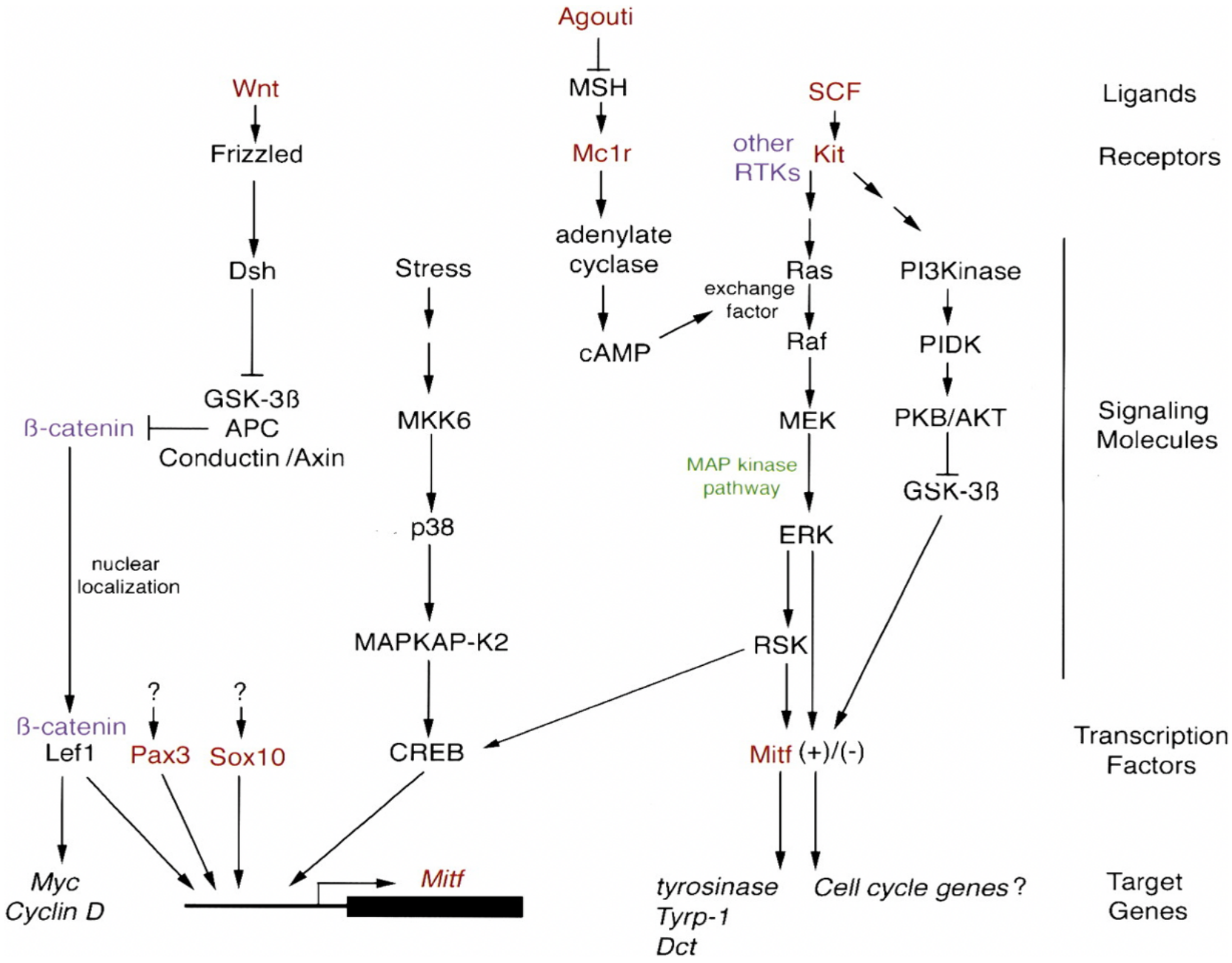
D. Blei, A. Ng, M. Jordan. *Journal of Machine Learning Research*, 3: 993-1022, 2003

cisTopic: LDA Model



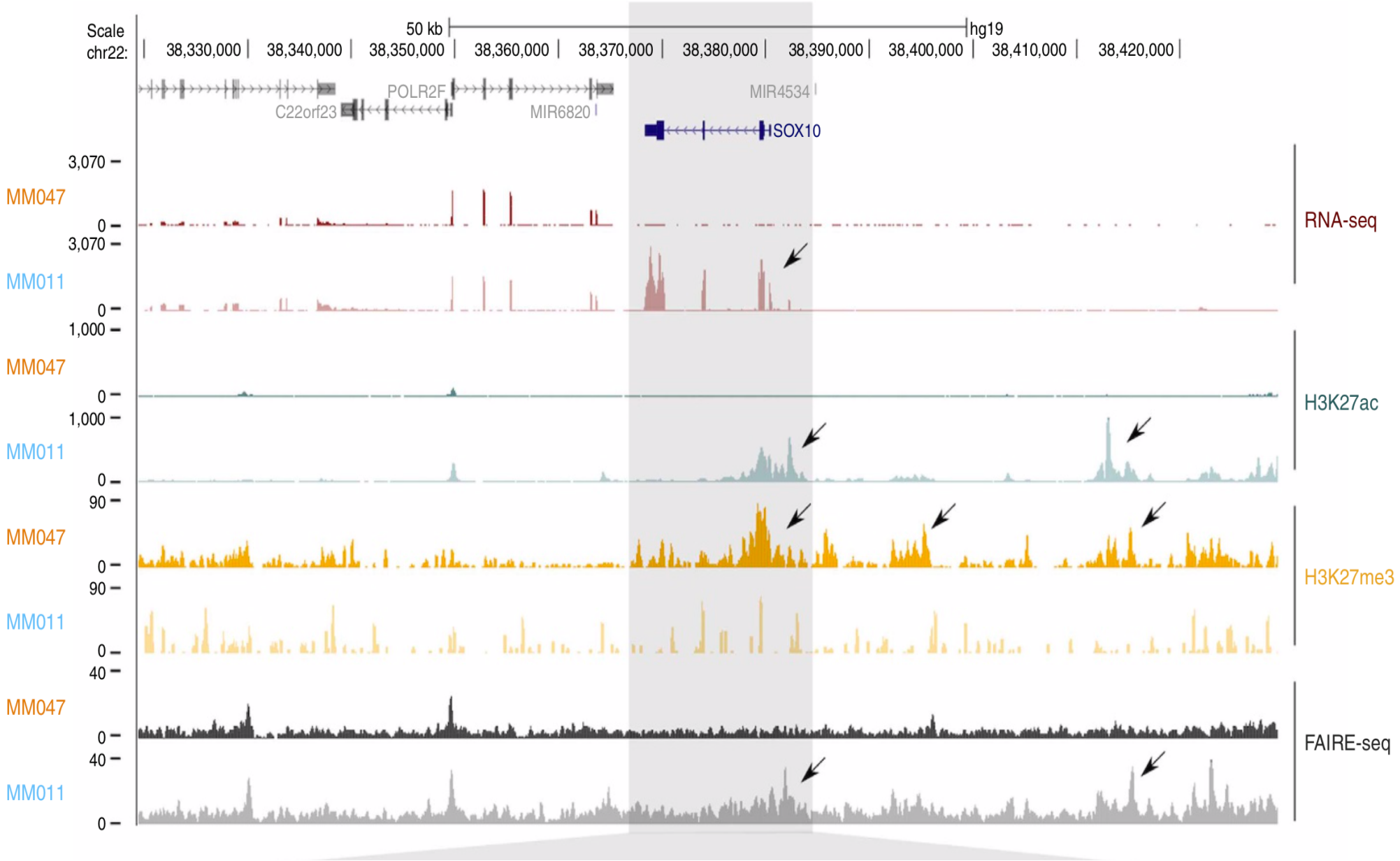
González-Blas et al. 2019 Nature Methods 16:397–400

Regulatory Pathways in Melanocyte Lineage



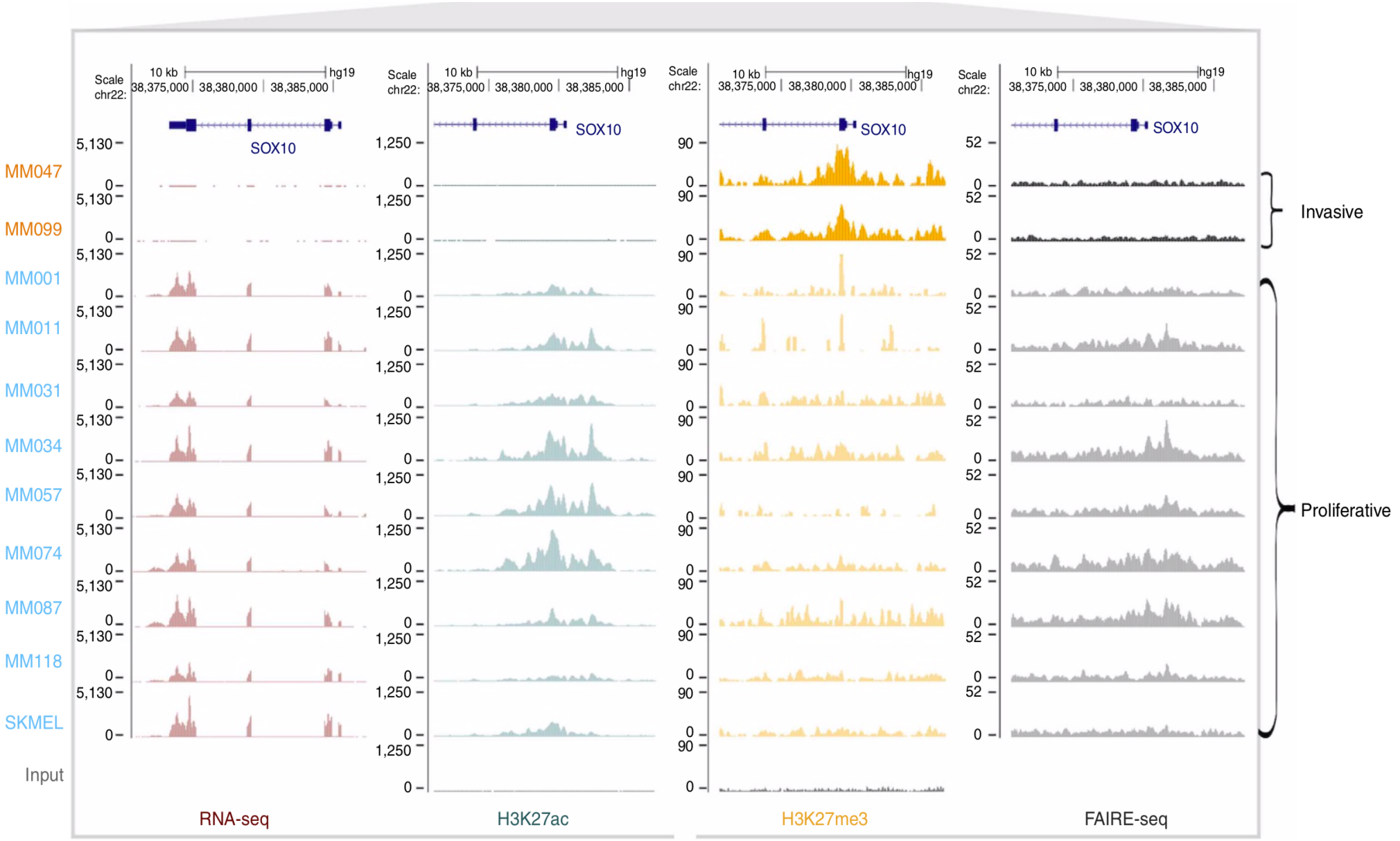
Colin R. Goding Genes Dev. 2000;14:1712-1728

Regulatory Landscape of Melanoma



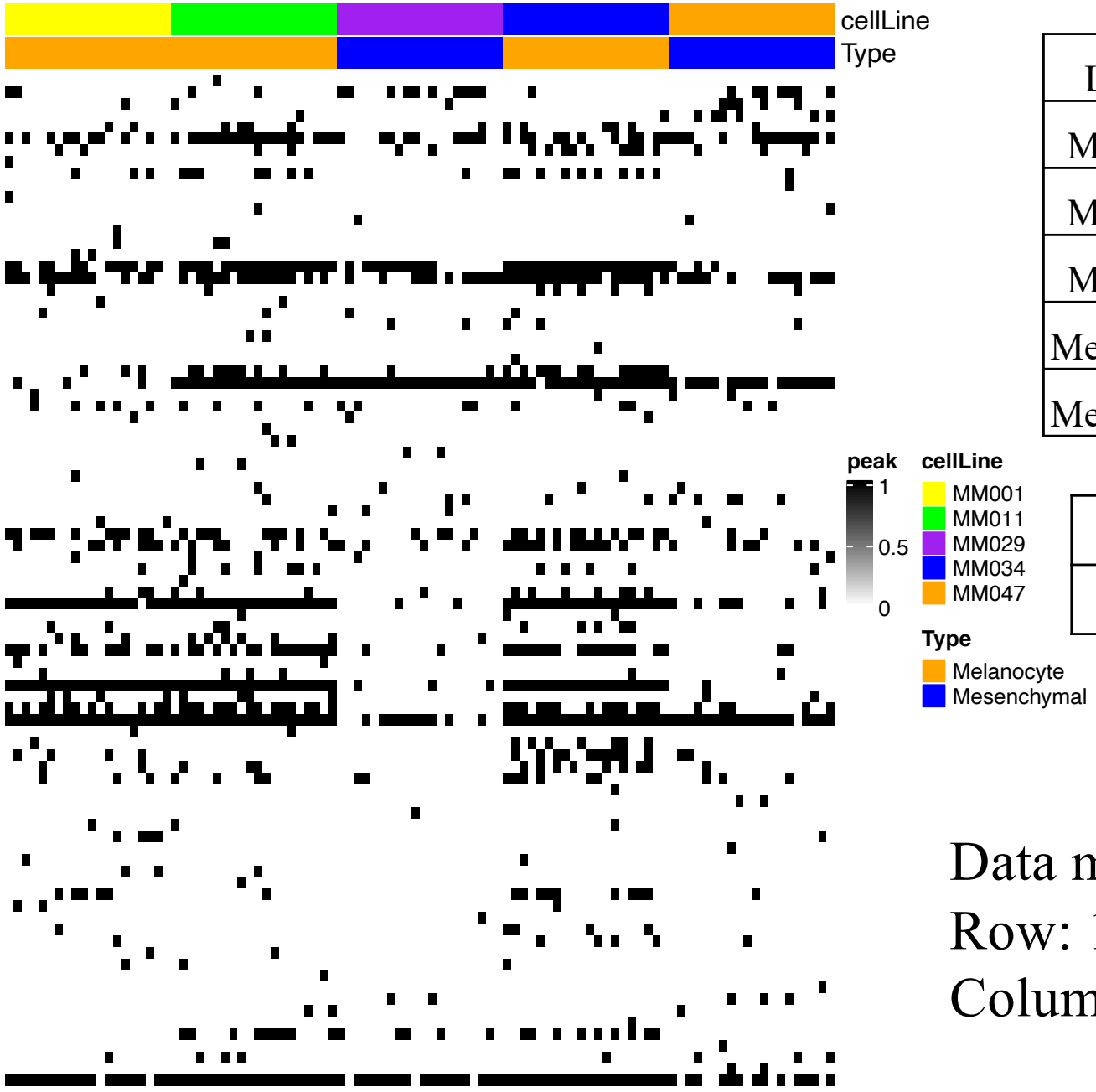
Verfaillie et al Nat Commun. 2015, 6:6683

Regulatory Landscape of Melanoma



Verfaillie et al Nat Commun. 2015, 6:6683

Heatmap of H3K27Ac ChIP-Seq Data

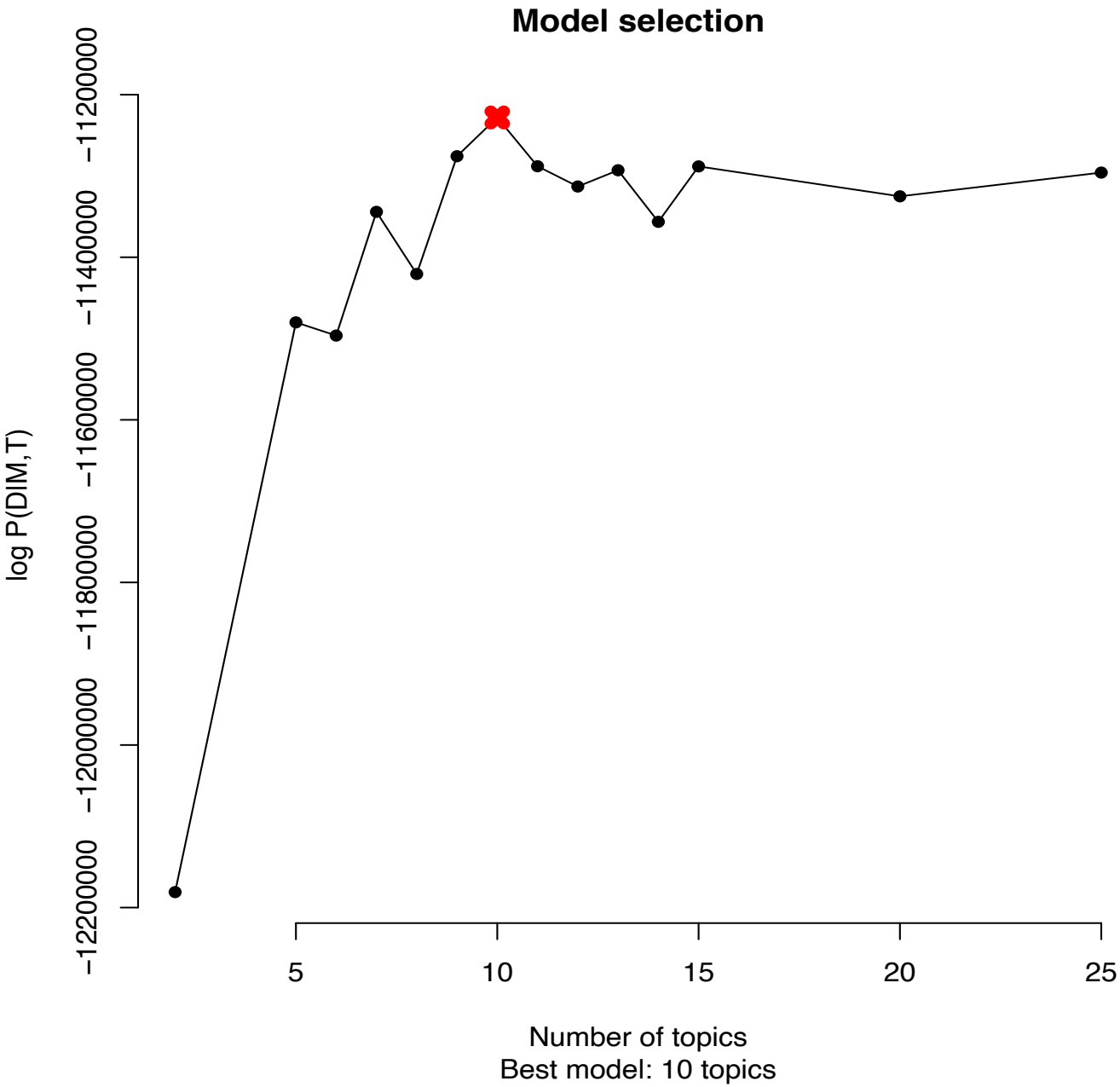


LineType	cellLine	freq
Melanocyte	MM001	20
Melanocyte	MM011	20
Melanocyte	MM034	20
Mesenchymal	MM029	20
Mesenchymal	MM047	20

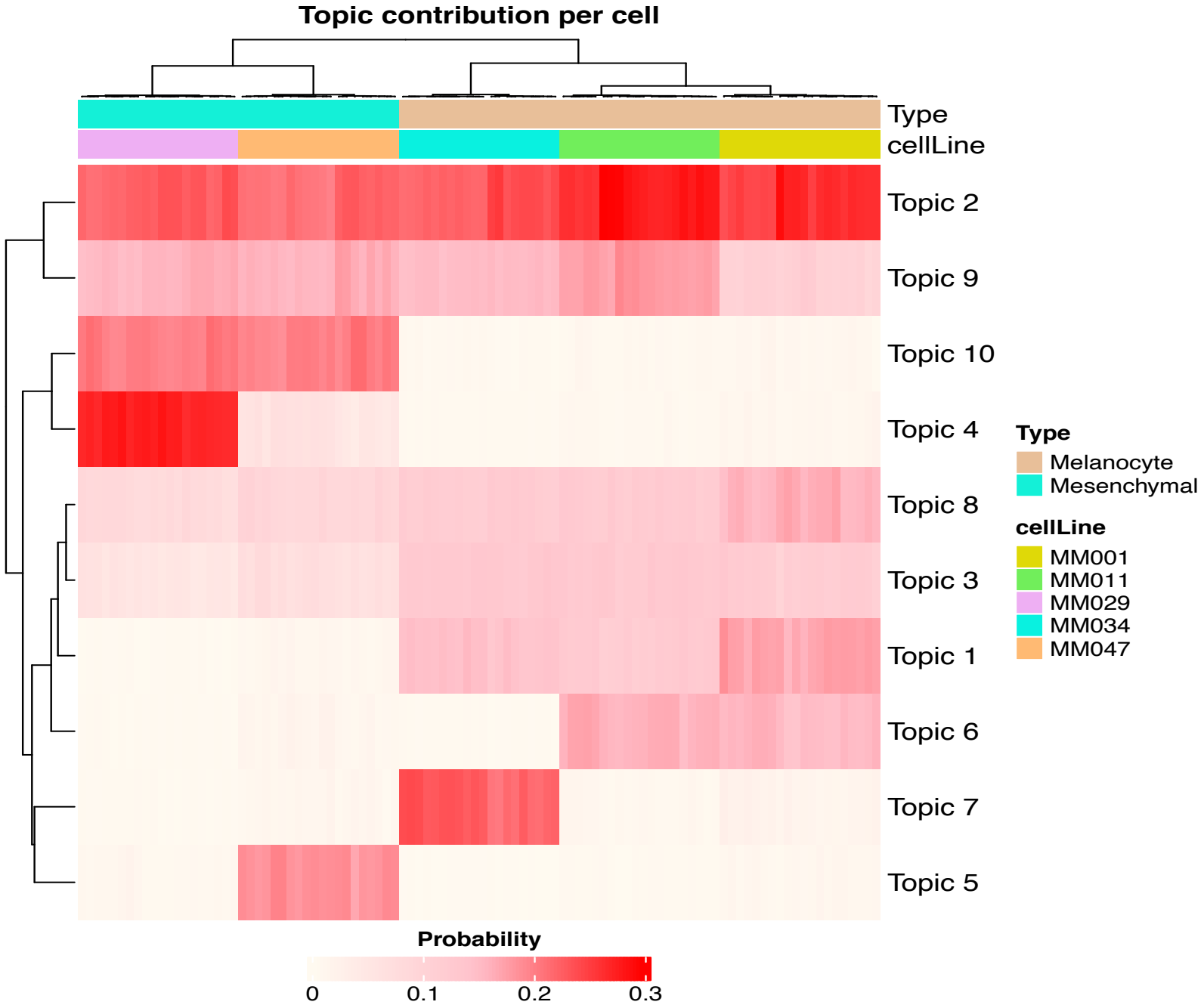
Proliferative	Melanocyte
Invasive	Mesenchymal

Data matrix
Row: 1135772 peaks
Column: 100 samples

Model Selection from Number of Topics



Heatmap of Topic Probability Distribution



TF Binding Sites of Topic 6

Rank	Motif	Name	P-value	log P-pvalue	q-value (Benjamini)	# Target Sequences with Motif	% of Targets Sequences with Motif
1		MITF(bHLH)/MastCells-MITF-ChIP-Seq(GSE48085)/Homer	1e-219	-5.059e+02	0.0000	1219.0	62.67%
2		NPAS(bHLH)/Liver-NPAS-ChIP-Seq(GSE39860)/Homer	1e-171	-3.944e+02	0.0000	1456.0	74.86%
3		BMAL1(bHLH)/Liver-Bmal1-ChIP-Seq(GSE39860)/Homer	1e-165	-3.808e+02	0.0000	1517.0	77.99%
4		MNT(bHLH)/HepG2-MNT-ChIP-Seq(Encode)/Homer	1e-162	-3.734e+02	0.0000	1163.0	59.79%
5		CLOCK(bHLH)/Liver-Clock-ChIP-Seq(GSE39860)/Homer	1e-158	-3.642e+02	0.0000	887.0	45.60%
6		USF1(bHLH)/GM12878-Usf1-ChIP-Seq(GSE32465)/Homer	1e-156	-3.613e+02	0.0000	814.0	41.85%
7		Max(bHLH)/K562-Max-ChIP-Seq(GSE31477)/Homer	1e-153	-3.534e+02	0.0000	955.0	49.10%
8		NPAS2(bHLH)/Liver-NPAS2-ChIP-Seq(GSE39860)/Homer	1e-150	-3.471e+02	0.0000	1245.0	64.01%
9		n-Myc(bHLH)/mES-nMyc-ChIP-Seq(GSE11431)/Homer	1e-149	-3.435e+02	0.0000	1020.0	52.44%
10		bHLHE41(bHLH)/proB-Bhlhe41-ChIP-Seq(GSE93764)/Homer	1e-145	-3.348e+02	0.0000	1321.0	67.92%

TF binding sites analyzed with homer

TF Binding Sites of Topic 7

Rank	Motif	Name	P-value	log P-pvalue	q-value (Benjamini)
1		MITF(bHLH)/MastCells-MITF-ChIP-Seq(GSE48085)/Homer	1e-168	-3.887e+02	0.0000
2		Sox9(HMG)/Limb-SOX9-ChIP-Seq(GSE73225)/Homer	1e-162	-3.733e+02	0.0000
3		Atf3(bZIP)/GBM-ATF3-ChIP-Seq(GSE33912)/Homer	1e-159	-3.669e+02	0.0000
4		Fra2(bZIP)/Striatum-Fra2-ChIP-Seq(GSE43429)/Homer	1e-157	-3.616e+02	0.0000
5		Fra1(bZIP)/BT549-Fra1-ChIP-Seq(GSE46166)/Homer	1e-156	-3.593e+02	0.0000
6		BMAL1(bHLH)/Liver-Bmal1-ChIP-Seq(GSE39860)/Homer	1e-155	-3.587e+02	0.0000
7		JunB(bZIP)/DendriticCells-Junb-ChIP-Seq(GSE36099)/Homer	1e-155	-3.572e+02	0.0000
8		Fos(bZIP)/TSC-Fos-ChIP-Seq(GSE110950)/Homer	1e-154	-3.555e+02	0.0000
9		NPAS(bHLH)/Liver-NPAS-ChIP-Seq(GSE39860)/Homer	1e-153	-3.542e+02	0.0000
10		AP-1(bZIP)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer	1e-152	-3.511e+02	0.0000

TF binding sites analyzed with homer

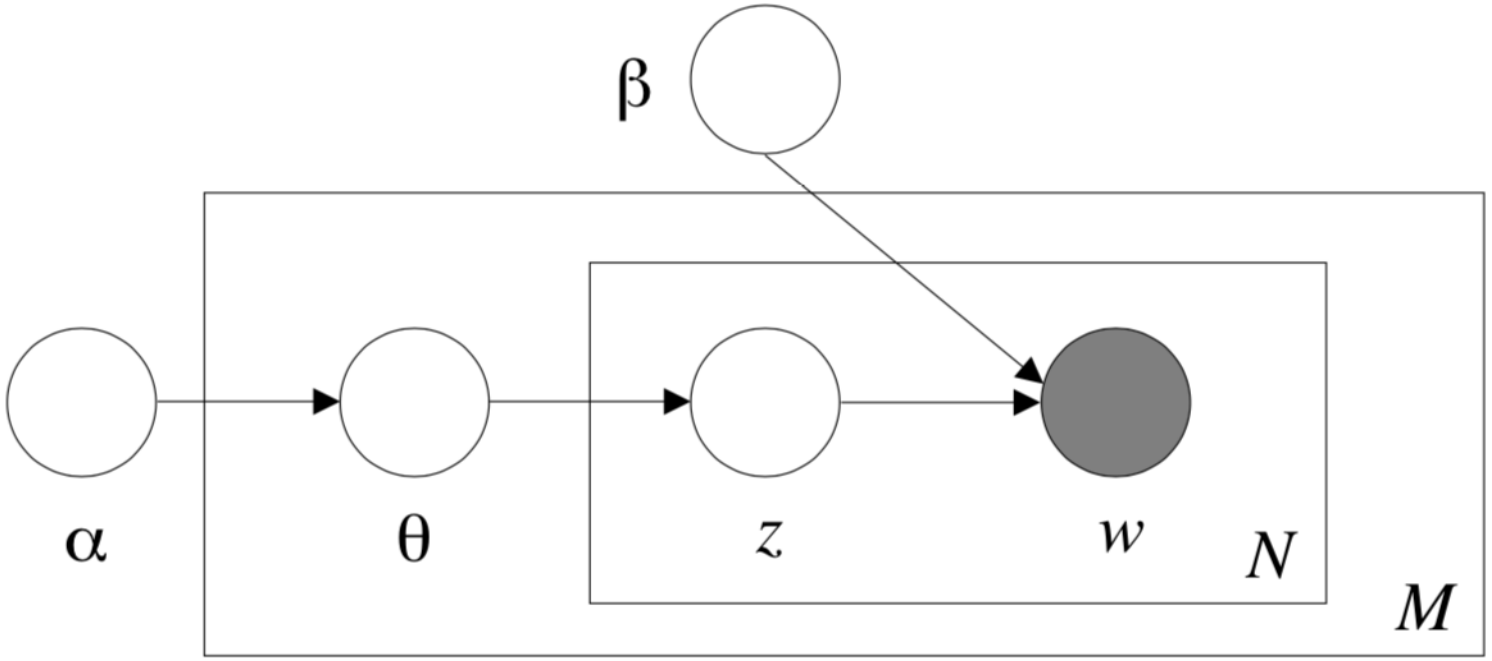
TF Binding Sites of Topic 10

Rank	Motif	Name	P-value	log P-pvalue	q-value (Benjamini)
1		Fra2(bZIP)/Striatum-Fra2-ChIP-Seq(GSE43429)/Homer	1e-624	-1.438e+03	0.0000
2		Fosl2(bZIP)/3T3L1-Fosl2-ChIP-Seq(GSE56872)/Homer	1e-613	-1.412e+03	0.0000
3		Fos(bZIP)/TSC-Fos-ChIP-Seq(GSE110950)/Homer	1e-606	-1.396e+03	0.0000
4		Fra1(bZIP)/BT549-Fra1-ChIP-Seq(GSE46166)/Homer	1e-601	-1.385e+03	0.0000
5		Atf3(bZIP)/GBM-ATF3-ChIP-Seq(GSE33912)/Homer	1e-598	-1.379e+03	0.0000
6		JunB(bZIP)/DendriticCells-Junb-ChIP-Seq(GSE36099)/Homer	1e-594	-1.369e+03	0.0000
7		BATF(bZIP)/Th17-BATF-ChIP-Seq(GSE39756)/Homer	1e-586	-1.350e+03	0.0000
8		Jun-AP1(bZIP)/K562-cJun-ChIP-Seq(GSE31477)/Homer	1e-565	-1.302e+03	0.0000
9		AP-1(bZIP)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer	1e-563	-1.297e+03	0.0000
10		Bach2(bZIP)/OCILy7-Bach2-ChIP-Seq(GSE44420)/Homer	1e-359	-8.277e+02	0.0000

TF binding sites analyzed with homer

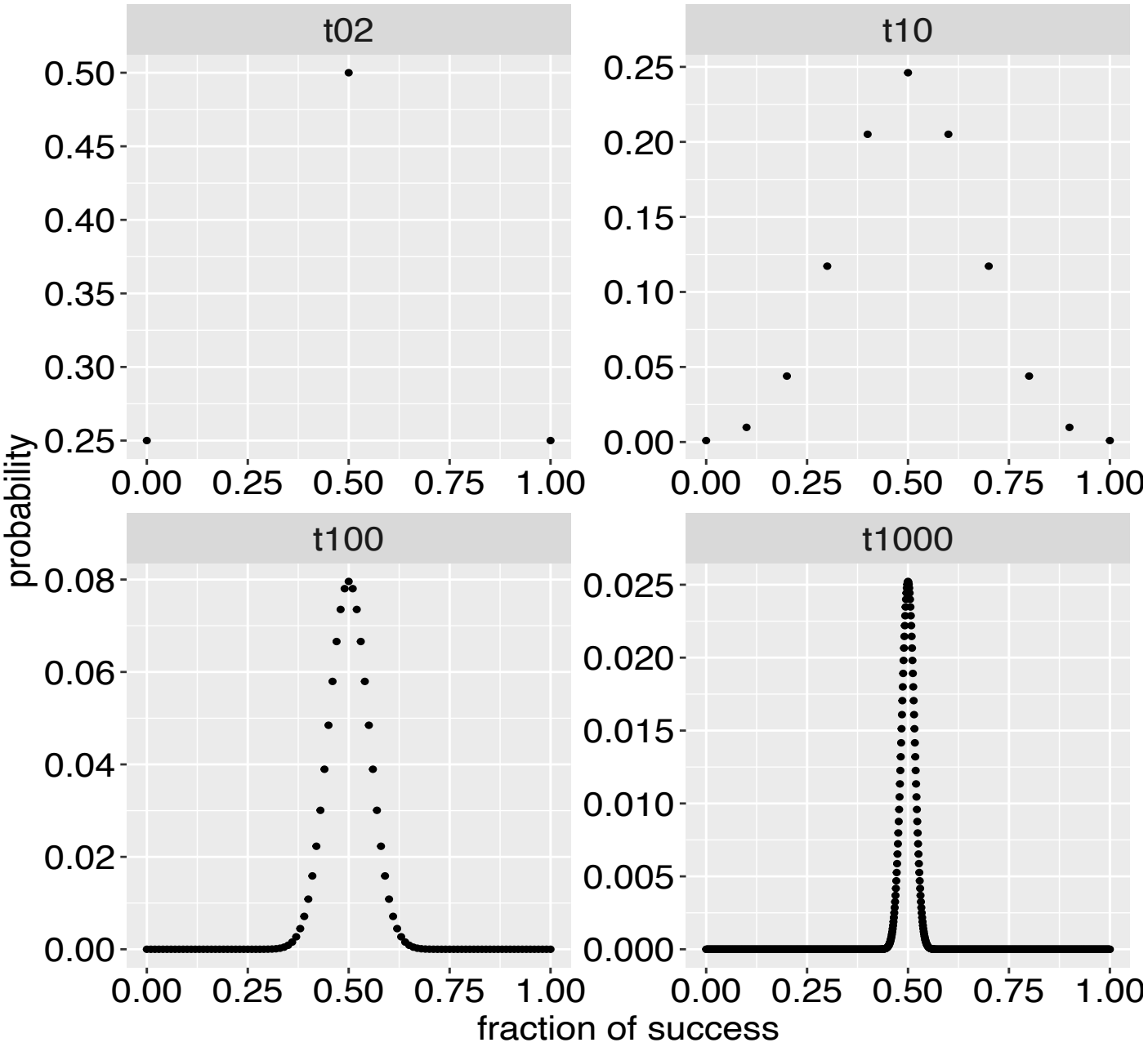
Algorithm of Latent Dirichlet Allocation (LDA)

LDA is a Bayesian Hierarchical Model and a Generative Statistical Model



D. Blei, A. Ng, M. Jordan. *Journal of Machine Learning Research*, 3: 993-1022, 2003

Traditional Statistics: Frequentist Approach



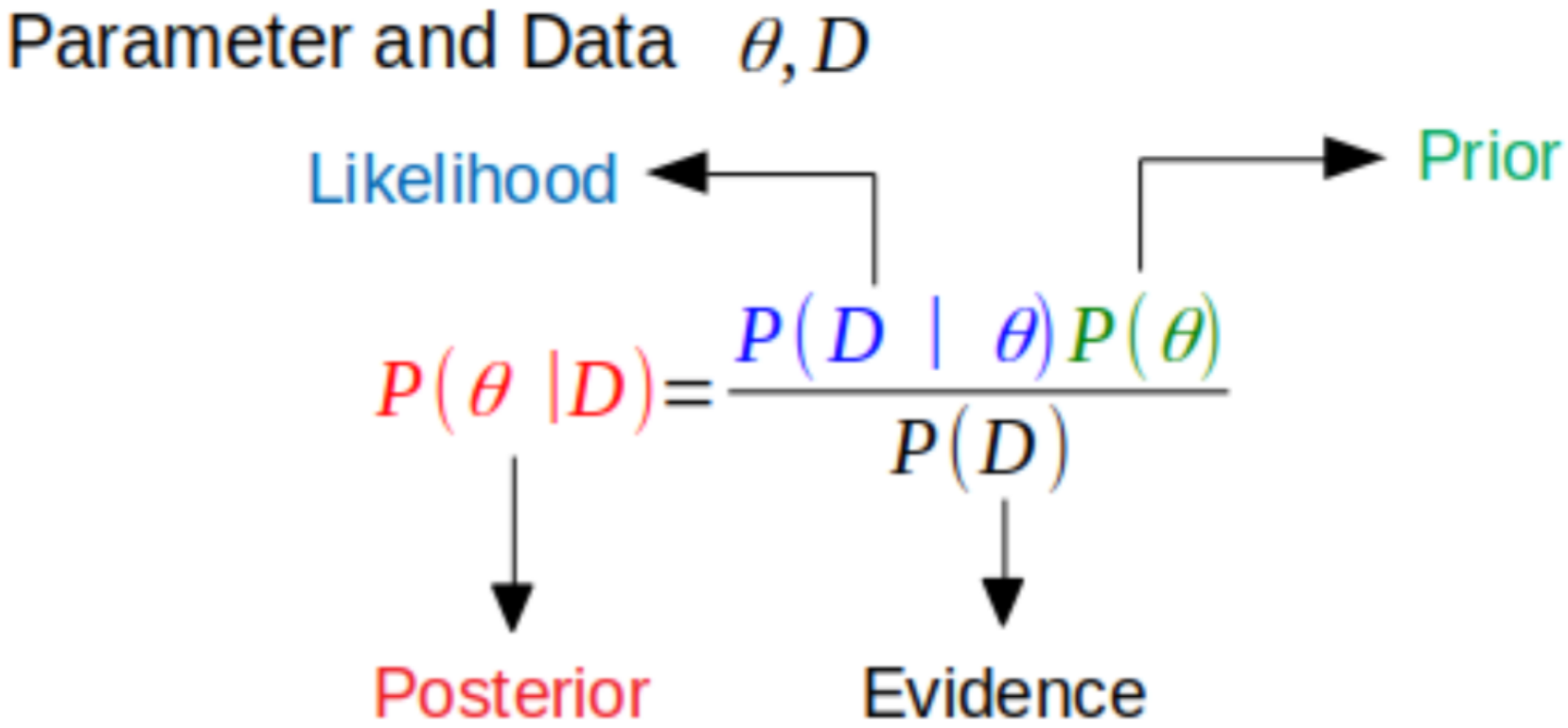
large sample size
a point estimate

Bayes' Theorem

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Thomas Bayes in 1763

Bayesian Analysis



Bayesian Analysis for Binary Variable

Bernoulli Likelihood $\longrightarrow$ $p(x \mid \theta) = \theta^{N_1}(1-\theta)^{N_0}$

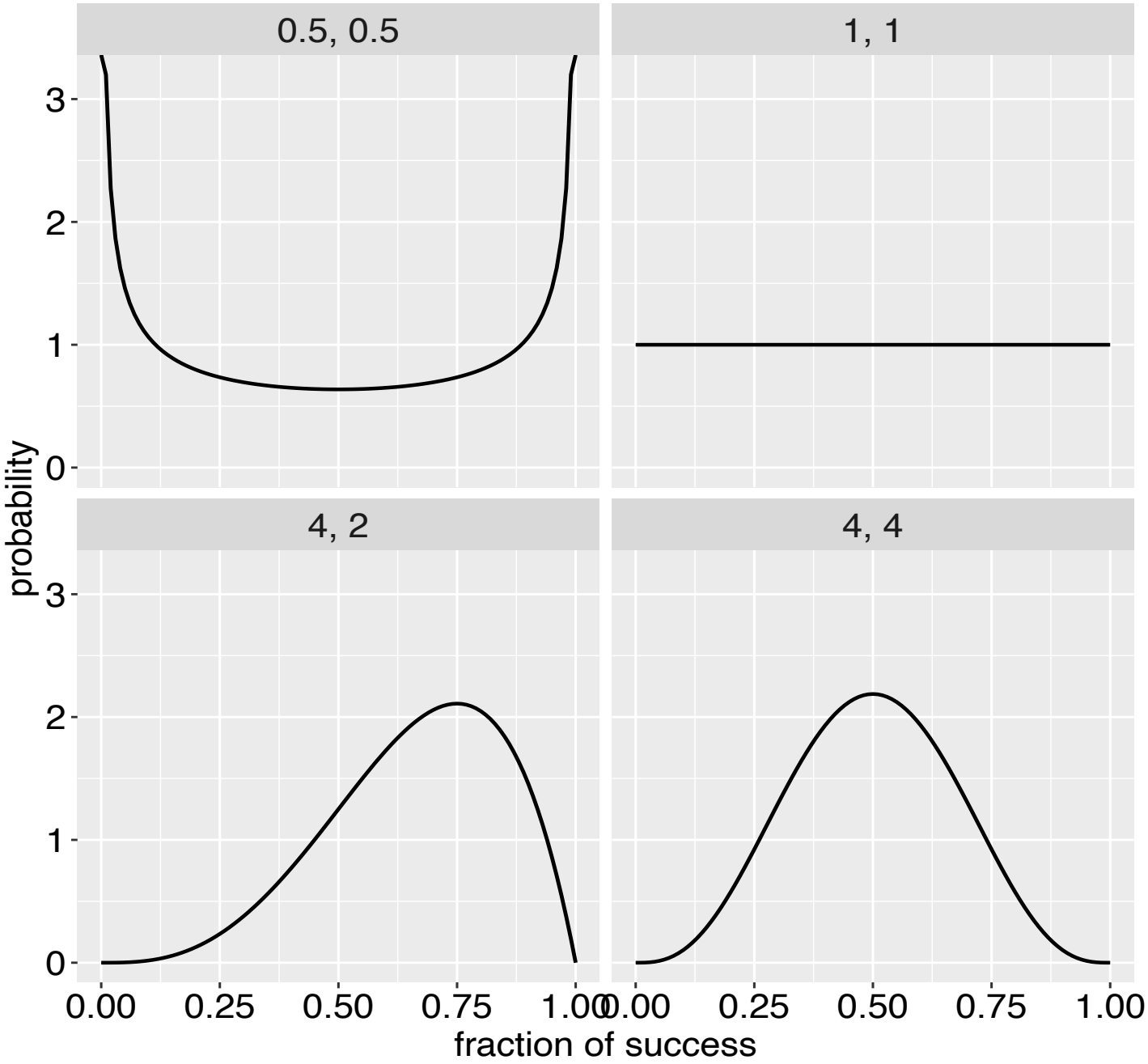
Beta Prior $\longrightarrow$ $p(\theta) = B(\theta \mid a,b) \propto \theta^{a-1}(1-\theta)^{b-1}$

$$p(\theta \mid x) \propto \theta^{N_1}(1-\theta)^{N_0} \times \theta^{a-1}(1-\theta)^{b-1}$$

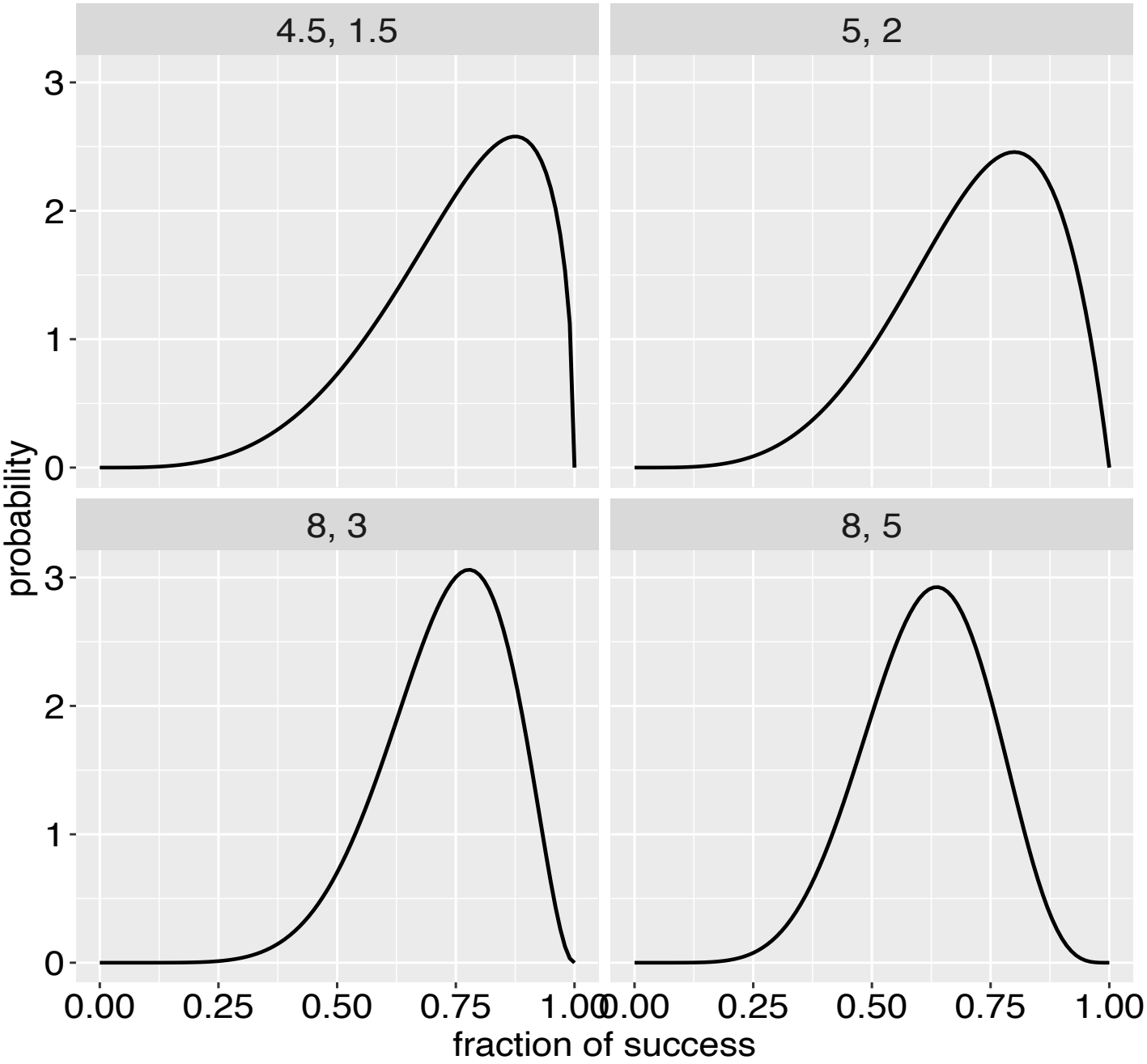
$\downarrow$

Beta Posterior $\longrightarrow$ $p(\theta \mid x) \propto B(N_1+a, N_0+b)$

Beta Prior

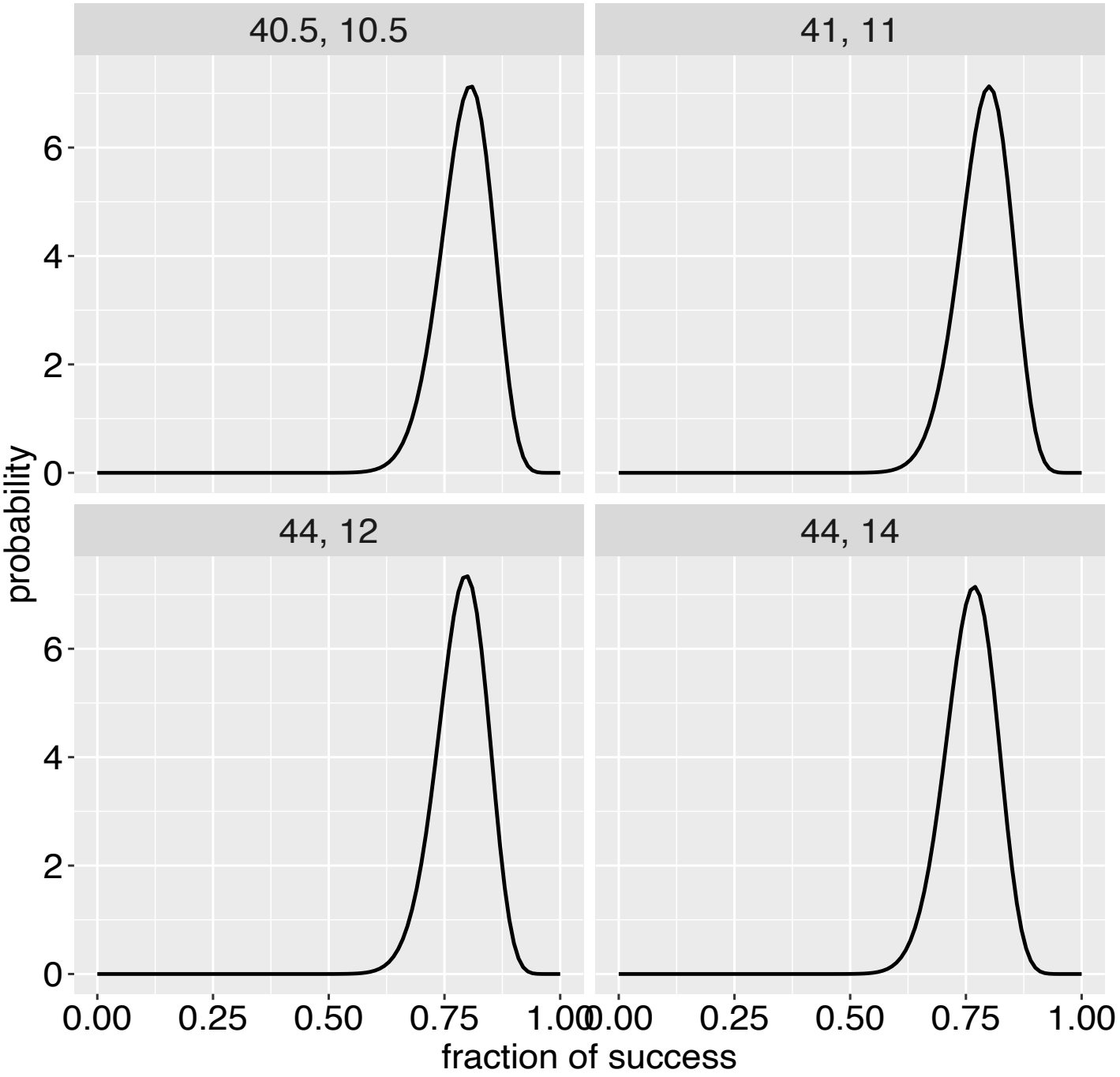

0.5, 0.5
1, 1
4, 2
4, 4
probability
0
1
2
3
0.00
0.25
0.50
0.75
1.00
fraction of success

Beta Posterior



D = (4, 1)

Beta Posterior



D = (40, 10)

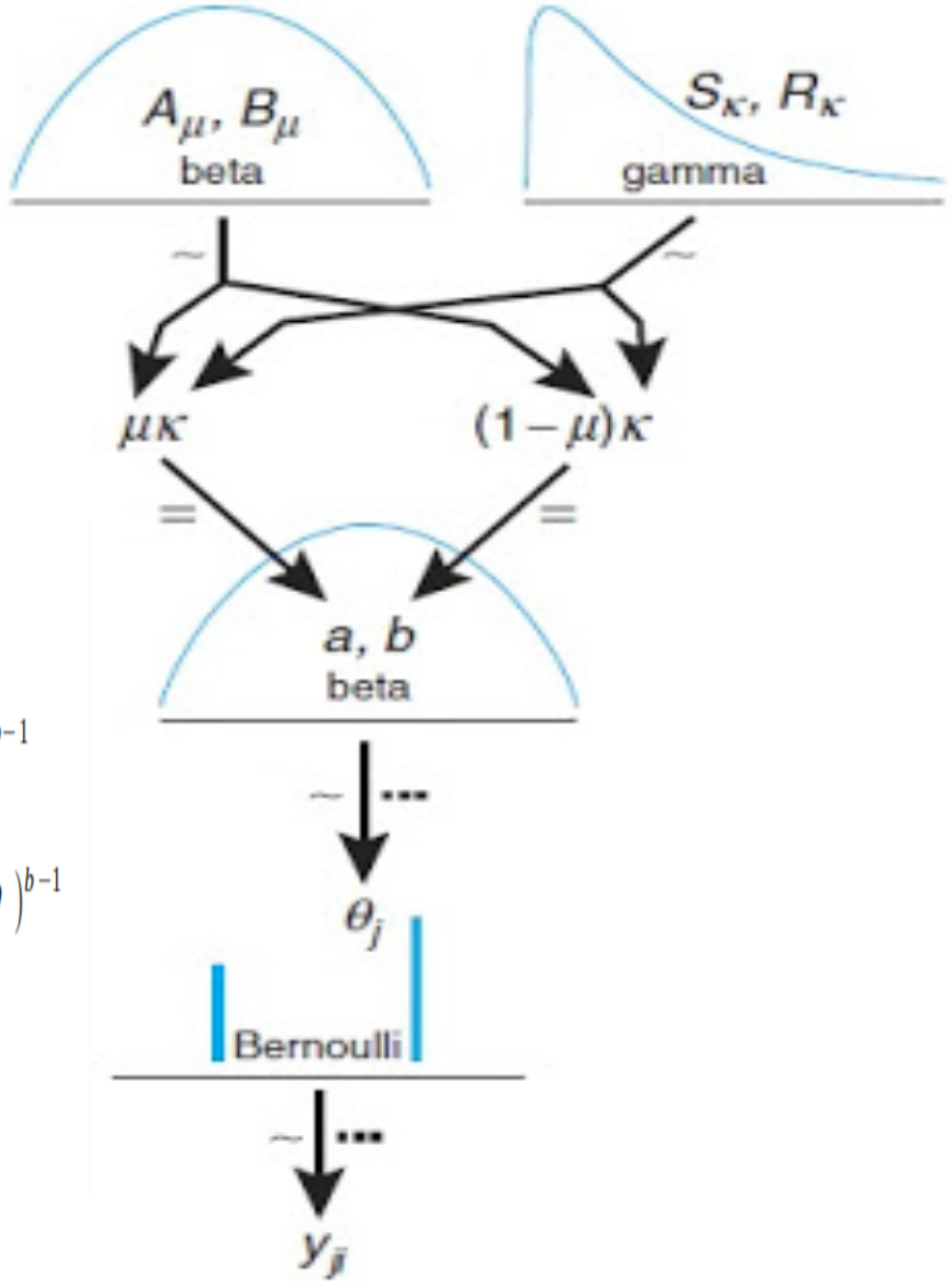
Bayesian Hierarchical Model

Bernoulli Likelihood $\longrightarrow$ $p(\theta \mid x) = \theta^{N_1}(1-\theta)^{N_0}$

Beta Prior $\longrightarrow$ $p(\theta) = B(\theta \mid a,b) \propto \theta^{a-1}(1-\theta)^{b-1}$

$$p(\theta \mid x) \propto \theta^{N_1}(1-\theta)^{N_0} \times \theta^{a-1}(1-\theta)^{b-1}$$

Beta Posterior $\longrightarrow$ $p(\theta \mid x) \propto B(N_1+a, N_0+b)$



Multinomial Distribution

$$f(x_1,\ldots,x_k;n,p_1,\ldots,p_k) = \Pr(X_1 = x_1 \text{ and } \ldots \text{ and } X_k = x_k)$$

$$= \begin{cases} \dfrac{n!}{x_1!\cdots x_k!} p_1^{x_1} \times \cdots \times p_k^{x_k}, & \text{when } \sum_{i=1}^k x_i = n \\ \\ 0 & \text{otherwise,} \end{cases}$$

$$\Gamma(n) = 1 \cdot 2 \cdot 3 \cdots (n-1) = (n-1)!$$

$$f(x_1,\ldots,x_k;p_1,\ldots,p_k) = \frac{\Gamma(\sum_i x_i + 1)}{\prod_i \Gamma(x_i+1)} \prod_{i=1}^k p_i^{x_i}$$

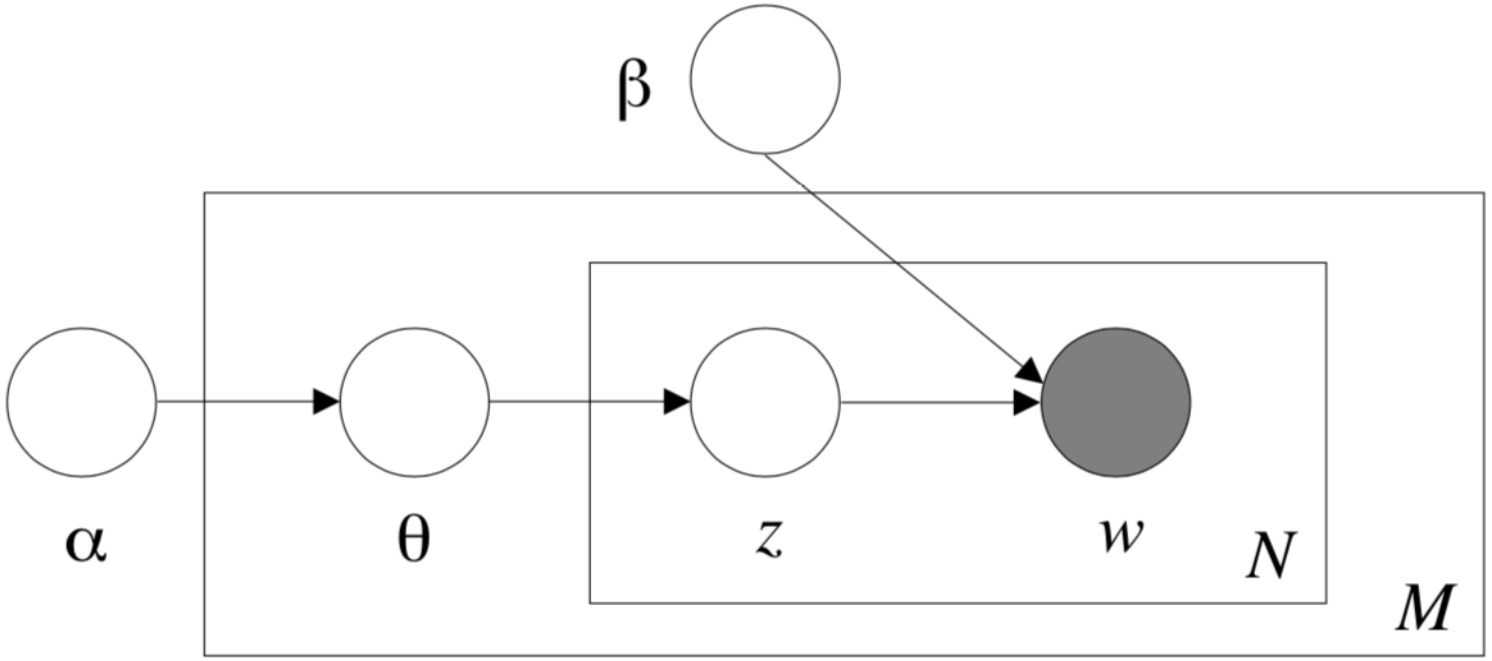
Dirichlet Distribution

$$p(\theta \mid \alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}$$

$$\theta_i \geq 0, \sum_{i=1}^{k} \theta_i = 1$$

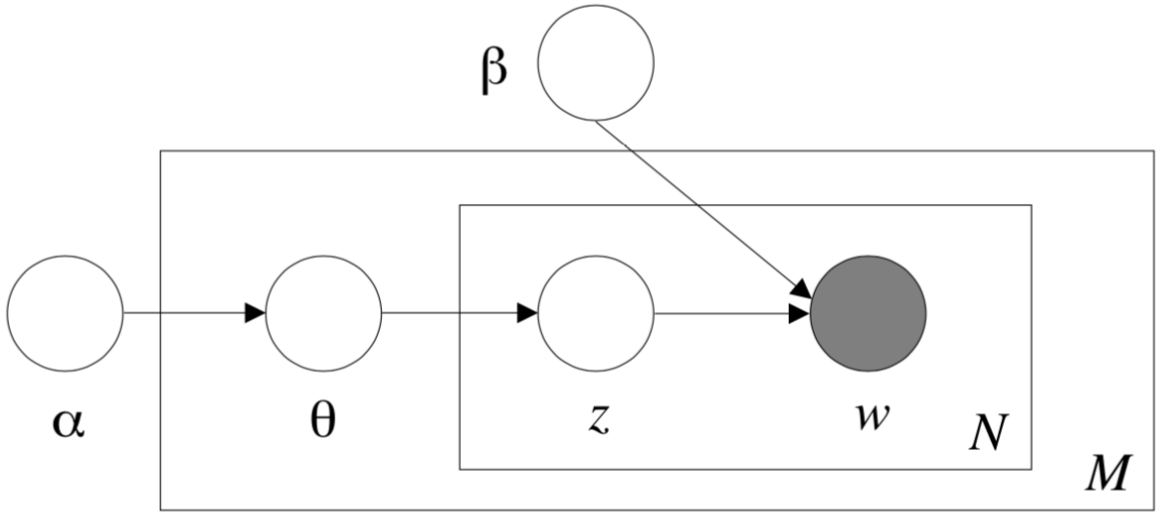
Algorithm of Latent Dirichlet Allocation (LDA)

LDA is a Bayesian Hierarchical Model and a Generative Statistical Model



D. Blei, A. Ng, M. Jordan. *Journal of Machine Learning Research*, 3: 993-1022, 2003

Algorithm of Latent Dirichlet Allocation (LDA)

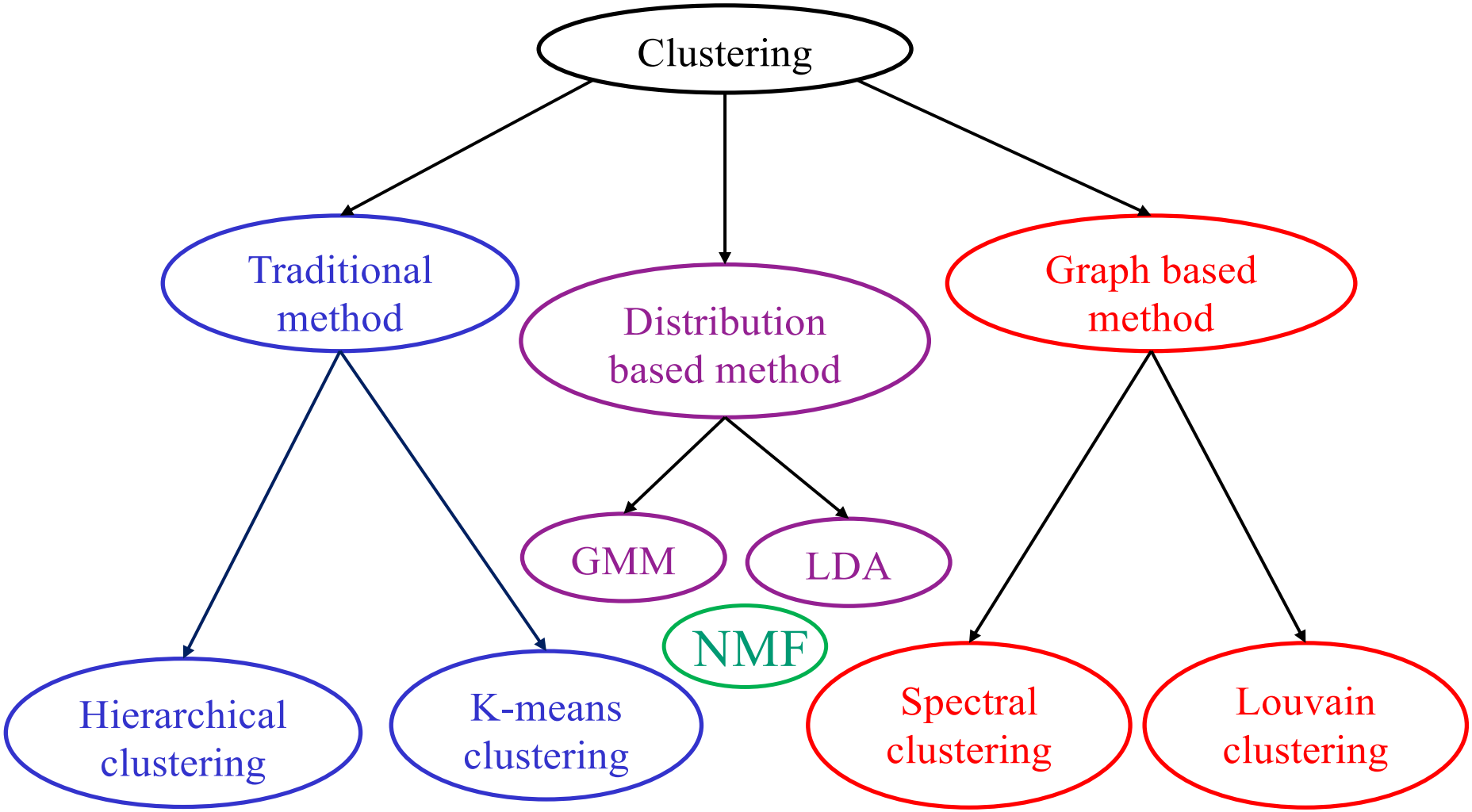


$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta)$$

For each document, update parameters till converge

D. Blei, A. Ng, M. Jordan. *Journal of Machine Learning Research*, 3: 993-1022, 2003

Outline of Clustering Methods



GMM: Gaussian Mixture Model
LDA: Latent Dirichlet Allocation
NMF: Non-negative matrix factorization

Contributed by Emily Tai