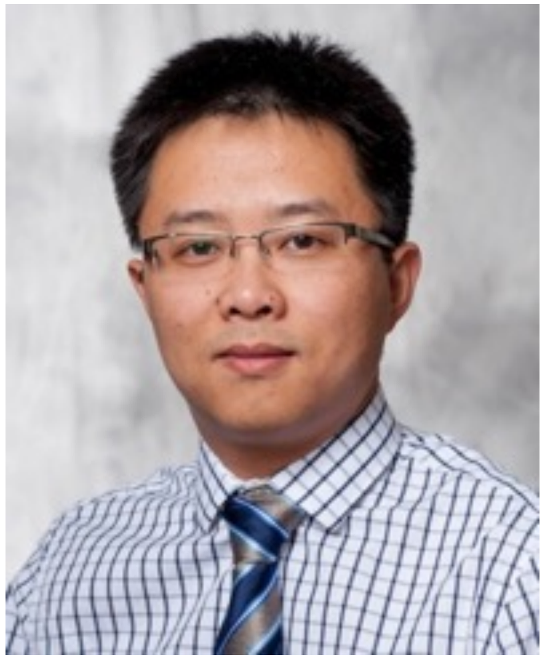
Beta Tucker Decomposition for DNA Methylation Data

Aaron Schein
UMass Amherst

Joint work with:

Pat Flaherty — UMass Amherst
Mingyuan Zhou — Univ. Texas at Austin
Dan Sheldon — UMass Amherst
Hanna Wallach — Microsoft Research

DNA methylation

CATTC**CG**CCTTCTCTCC**CG**AGG

DNA methylation

CATTC**CG**CCTTCTCTCCC**CG**AGG

CpG dinucleotides

DNA methylation

M

CATTC**CG**CCTTCTCTCCC**CG**AGG

methylated

unmethylated

DNA methylation

CGAGGCATTC**CG**CCTTCTCTCC**CG**AGGCATTC**CG**CCT
T**CG**A**CGCG**CCTTCTCTCC**CGCGCG**A**CGCG**CCTTCTCT
CC**CGCGCG**A**CGCG**CCTTCTCTCC**CGCG**CT**CG**A**CGCG**
CCTTCTCTCC**CGCGCG**A**CGCG**CCTTCTCTCC**CGCGCG**
A**CGCG**CCTTCTCTCC**CGCG**C**CG**A**CGCG**CCTTCTCTCC
CGCGCGA**CGCG**CCTTCTCTCC**CGCGCG**A**CGCG**CCTT
CTCTCC**CGCG**TCC**CGCG**A**CGCG**CCTTCTCTCC**CGCG**A
GGCATTC**CG**CCTTCTTTTTTTTTTTTT**CG**A**CGCG**CCTTCT
CTCC**CGCGCG**A**CGCG**CCTTCTCTCC**CGCG**TTTTTCTC
C**CG**AGGCATTC**CG**CCTTCTC**CG**A**CGCG**CCTTCTCTCC
CGCGTTCTCTAG**CG**CCTTCTCTCC**CG**A**CG**A**CGCG**CCT
TCTCTCC**CGCGCG**A**CGCG**A**CGCG**CCTTCTCTCC**CGC**
GCGCCTTCTCTCC**CGCG**CCTTCTCTCC**CG**A**CG**CCTTC
TCTCC**CG**A**CG**CCTTCTCTCC**CG**A**CGCG**CCTTCTCTCC
CGCGCCTTCTCTCC**CGCG**CCTTCTCTCC**CG**A**CG**CCTT

CATTC**CG**CCTTCTGCTCTCTAGTCCCCCAGGCTGGAT
TGCTACACCTTCTCTAGTCCCCCAGGCTGGATTGCTAC
ACCTATCTCC**CG**AGGCATTC**CG**CCTTCTCTCC**CG**AGG
CATTC**CG**CCTTCTCTCC**CG**AGGCATTC**CG**CCTTCTTTT
TTTTTTTTTTTTTCTCC**CG**AGGCATTC**CG**CCTTCTCTTCT
CTAGTCCCCCAGGCTGGATTGCTACACCTTCTCTAGTC
CCCCAGGCTGGATTGCTACACCTTCTCTAGTCCCCCA
GGCTGGATTGCTACACCTCC**CG**AGGCATGCATTC**CG**
CCTTTCTCTAGTCCCCCAGGCTGGATTGCTACACCTTC
TCTAGTCCCCCAGGCTGGATTGCTACACCTCTCTC**CG**
AGGCATTC**CG**CCTTCCTCTCCTCTCTCTCC**CG**AGTCTC
TAGTCCCCCAGGCTGGATTGCTACACCTTCTCTAGTCC
CCCAGGCTGGATTGCTACACCTGCATTC**CG**CCTTCTC
TTTTTCC**CG**AGGCATTTCTCTAGTCCCCCAGGCTGGAT
TGCTACACCTTCTCTAGTCCCCCAGGCTGGATTGCTAC

Gene

DNA methylation

CGAGGCATTCCGCCTTCTCTCCCGAGGCATTCCGCCT
TCGACGCGCCTTCTCTCCCGCGCGACGCGCCTTCTCT
CCCGCGCGACGCGCCTTCTCTCCCGCGCTCGACGCG
CCTTCTCTCCCGCGCGACGCGCCTTCTCTCCCGCGCG
ACGCGCCTTCTCTCCCGCGCCGACGCGCCTTCTCTCC
CGCGCGACGCGCCTTCTCTCCCGCGCGACGCGCCTT
CTCTCCCGCGTCCCGCGACGCGCCTTCTCTCCCGCGA
GGCATTCCGCCTTCTTTTTTTTTTTTTCGACGCGCCTTCT
CTCCCGCGCGACGCGCCTTCTCTCCCGCGTTTTTCTC
CCGAGGCATTCCGCCTTCTCCGACGCGCCTTCTCTCC
CGCGTTCTCTAGCGCCTTCTCTCCCGACGACGCGCCT
TCTCTCCCGCGCGACGCGACGCGCCTTCTCTCCCGC
GCGCCTTCTCTCCCGCGCCTTCTCTCCCGACGCCTTC
TCTCCCGACGCCTTCTCTCCCGACGCGCCTTCTCTCC
CGCGCCTTCTCTCCCGCGCCTTCTCTCCCGACGCCTT

CpG island

CATTCCGCCTTCTGCTCTCTAGTCCCCCAGGCTGGAT
TGCTACACCTTCTCTAGTCCCCCAGGCTGGATTGCTAC
ACCTATCTCCCGAGGCATTCCGCCTTCTCTCCCGAGG
CATTCCGCCTTCTCTCCCGAGGCATTCCGCCTTCTTTT
TTTTTTTTTTTTTCTCCCGAGGCATTCCGCCTTCTCTTCT
CTAGTCCCCCAGGCTGGATTGCTACACCTTCTCTAGTC
CCCCAGGCTGGATTGCTACACCTTCTCTAGTCCCCCA
GGCTGGATTGCTACACCTCCCGAGGCATGCATTCCG
CCTTTCTCTAGTCCCCCAGGCTGGATTGCTACACCTTC
TCTAGTCCCCCAGGCTGGATTGCTACACCTCTCTCCG
AGGCATTCCGCCTTCCTCTCCTCTCTCTCCCGAGTCTC
TAGTCCCCCAGGCTGGATTGCTACACCTTCTCTAGTCC
CCCAGGCTGGATTGCTACACCTGCATTCCGCCTTCTC
TTTTTCCCGAGGCATTTCTCTAGTCCCCCAGGCTGGAT
TGCTACACCTTCTCTAGTCCCCCAGGCTGGATTGCTAC

DNA methylation

CGAGGCATTCCGCCTTCTCTCCCGAGGCATTCCGCCT
TCGACGCGCCTTCTCTCCCGCGCGACGCGCCTTCTCT
CCCGCGCGACGCGCCTTCTCTCCCGCGCTCGACGCG
CCTTCTCTCCCGCGCGACGCGCCTTCTCTCCCGCGCG
ACGCGCCTTCTCTCCCGCGCCGACGCGCCTTCTCTCC
CGCGCGACGCGCCTTCTCTCCCGCGCGACGCGCCTT
CTCTCCCGCGTCCCGCGACGCGCCTTCTCTCCCGCGA
GGCATTCCGCCTTCTTTTTTTTTTTTCGACGCGCCTTCT
CTCCCGCGCGACGCGCCTTCTCTCCCGCGTTTTTCTC
CCGAGGCATTCCGCCTTCTCCGACGCGCCTTCTCTCC
CGCGTTCTCTAGCGCCTTCTCTCCCGACGACGCGCCT
TCTCTCCCGCGCGACGCGACGCGCCTTCTCTCCCGC
GCGCCTTCTCTCCCGCGCCTTCTCTCCCGACGCCTTC
TCTCCCGACGCCTTCTCTCCCGACGCGCCTTCTCTCC
CGCGCCTTCTCTCCCGCGCCTTCTCTCCCGACGCCTT

CATTCCGCCTTCTGCTCTCTAGTCCCCCAGGCTGGAT
TGCTACACCTTCTCTAGTCCCCCAGGCTGGATTGCTAC
ACCTATCTCCCGAGGCATTCCGCCTTCTCTCCCGAGG
CATTCCGCCTTCTCTCCCGAGGCATTCCGCCTTCTTTT
TTTTTTTTTTTTTCTCCCGAGGCATTCCGCCTTCTCTTCT
CTAGTCCCCCAGGCTGGATTGCTACACCTTCTCTAGTC
CCCCAGGCTGGATTGCTACACCTTCTCTAGTCCCCCA
GGCTGGATTGCTACACCTCCCGAGGCATGCATTCCG
CCTTTCTCTAGTCCCCCAGGCTGGATTGCTACACCTTC
TCTAGTCCCCCAGGCTGGATTGCTACACCTCTCTCCG
AGGCATTCCGCCTTCCTCTCCTCTCTCTCCCGAGTCTC
TAGTCCCCCAGGCTGGATTGCTACACCTTCTCTAGTCC
CCCAGGCTGGATTGCTACACCTGCATTCCGCCTTCTC
TTTTTCCCGAGGCATTTCTCTAGTCCCCCAGGCTGGAT
TGCTACACCTTCTCTAGTCCCCCAGGCTGGATTGCTAC

CpG island (often in the promoter region)

DNA methylation

CGAGGCATTCCGCCTTCTCTCCCGAGGCATTCCGCCT
TCGACGCGCCTTCTCTCCCGCGCGACGCGCCTTCTCT
CCCGCGCGACGCGCCTTCTCTCCCGCGCTCGACGCG
CCTTCTCTCCCGCGCGACGCGCCTTCTCTCCCGCGCG
ACGCGCCTTCTCTCCCGCGCCGACGCGCCTTCTCTCC
CGCGCGACGCGCCTTCTCTCCCGCGCGACGCGCCTT
CTCTCCCGCGTCCCGCGACGCGCCTTCTCTCCCGCGA
GGCATTCCGCCTTCTTTTTTTTTTTTCGACGCGCCTTCT
CTCCCGCGCGACGCGCCTTCTCTCCCGCGTTTTTCTC
CCGAGGCATTCCGCCTTCTCCGACGCGCCTTCTCTCC
CGCGTTCTCTAGCGCCTTCTCTCCCGACGACGCGCCT
TCTCTCCCGCGCGACGCGACGCGCCTTCTCTCCCGC
GCGCCTTCTCTCCCGCGCCTTCTCTCCCGACGCCTTC
TCTCCCGACGCCTTCTCTCCCGACGCGCCTTCTCTCC
CGCGCCTTCTCTCCCGCGCCTTCTCTCCCGACGCCTT

CATTCCGCCTTCTGCTCTCTAGTCCCCCAGGCTGGAT
TGCTACACCTTCTCTAGTCCCCCAGGCTGGATTGCTAC
ACCTATCTCCCGAGGCATTCCGCCTTCTCTCCCGAGG
CATTCCGCCTTCTCTCCCGAGGCATTCCGCCTTCTTTT
TTTTTTTTTTTTTCTCCCGAGGCATTCCGCCTTCTCTTCT
CTAGTCCCCCAGGCTGGATTGCTACACCTTCTCTAGTC
CCCCAGGCTGGATTGCTACACCTTCTCTAGTCCCCCA
GGCTGGATTGCTACACCTCCCGAGGCATGCATTCCG
CCTTTCTCTAGTCCCCCAGGCTGGATTGCTACACCTTC
TCTAGTCCCCCAGGCTGGATTGCTACACCTCTCTCCG
AGGCATTCCGCCTTCCTCTCCTCTCTCTCCCGAGTCTC
TAGTCCCCCAGGCTGGATTGCTACACCTTCTCTAGTCC
CCCAGGCTGGATTGCTACACCTGCATTCCGCCTTCTC
TTTTTCCCGAGGCATTTCTCTAGTCCCCCAGGCTGGAT
TGCTACACCTTCTCTAGTCCCCCAGGCTGGATTGCTAC

Gene is **silenced**

DNA methylation

CGAGGCATTCCGCCTTCTCTCCCGAGGCATTCCGCCT
TCGACGCGCCTTCTCTCCCGCGCGACGCGCCTTCTCT
CCCGCGCGACGCGCCTTCTCTCCCGCGCTCGACGCG
CCTTCTCTCCCGCGCGACGCGCCTTCTCTCCCGCGCG
ACGCGCCTTCTCTCCCGCGCCGACGCGCCTTCTCTCC
CGCGCGACGCGCCTTCTCTCCCGCGCGACGCGCCTT
CTCTCCCGCGTCCCGCGACGCGCCTTCTCTCCCGCGA
GGCATTCCGCCTTCTTTTTTTTTTTTCGACGCGCCTTCT
CTCCCGCGCGACGCGCCTTCTCTCCCGCGTTTTTCTC
CCGAGGCATTCCGCCTTCTCCGACGCGCCTTCTCTCC
CGCGTTCTCTAGCGCCTTCTCTCCCGACGACGCGCCT
TCTCTCCCGCGCGACGCGACGCGCCTTCTCTCCCGC
GCGCCTTCTCTCCCGCGCCTTCTCTCCCGACGCCTTC
TCTCCCGACGCCTTCTCTCCCGACGCGCCTTCTCTCC
CGCGCCTTCTCTCCCGCGCCTTCTCTCCCGACGCCTT

CATTCCGCCTTCTGCTCTCTAGTCCCCCAGGCTGGAT
TGCTACACCTTCTCTAGTCCCCCAGGCTGGATTGCTAC
ACCTATCTCCCGAGGCATTCCGCCTTCTCTCCCGAGG
CATTCCGCCTTCTCTCCCGAGGCATTCCGCCTTCTTTT
TTTTTTTTTTTTCTCCCGAGGCATTCCGCCTTCTCTTCT
CTAGTCCCCCAGGCTGGATTGCTACACCTTCTCTAGTC
CCCCAGGCTGGATTGCTACACCTTCTCTAGTCCCCCA
GGCTGGATTGCTACACCTCCCGAGGCATGCATTCCG
CCTTTCTCTAGTCCCCCAGGCTGGATTGCTACACCTTC
TCTAGTCCCCCAGGCTGGATTGCTACACCTCTCTCCG
AGGCATTCCGCCTTCCTCTCCTCTCTCTCCCGAGTCTC
TAGTCCCCCAGGCTGGATTGCTACACCTTCTCTAGTCC
CCCAGGCTGGATTGCTACACCTGCATTCCGCCTTCTC
TTTTTCCCGAGGCATTTCTCTAGTCCCCCAGGCTGGAT
TGCTACACCTTCTCTAGTCCCCCAGGCTGGATTGCTAC

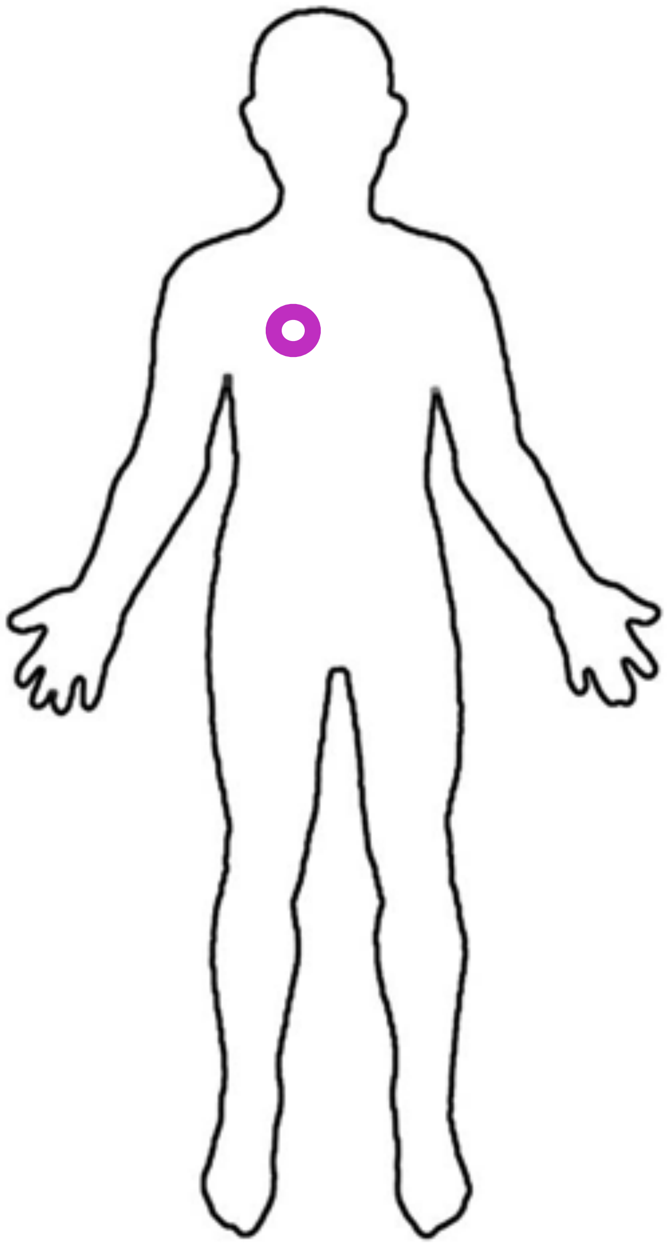
Gene is expressed

Abnormal DNA methylation

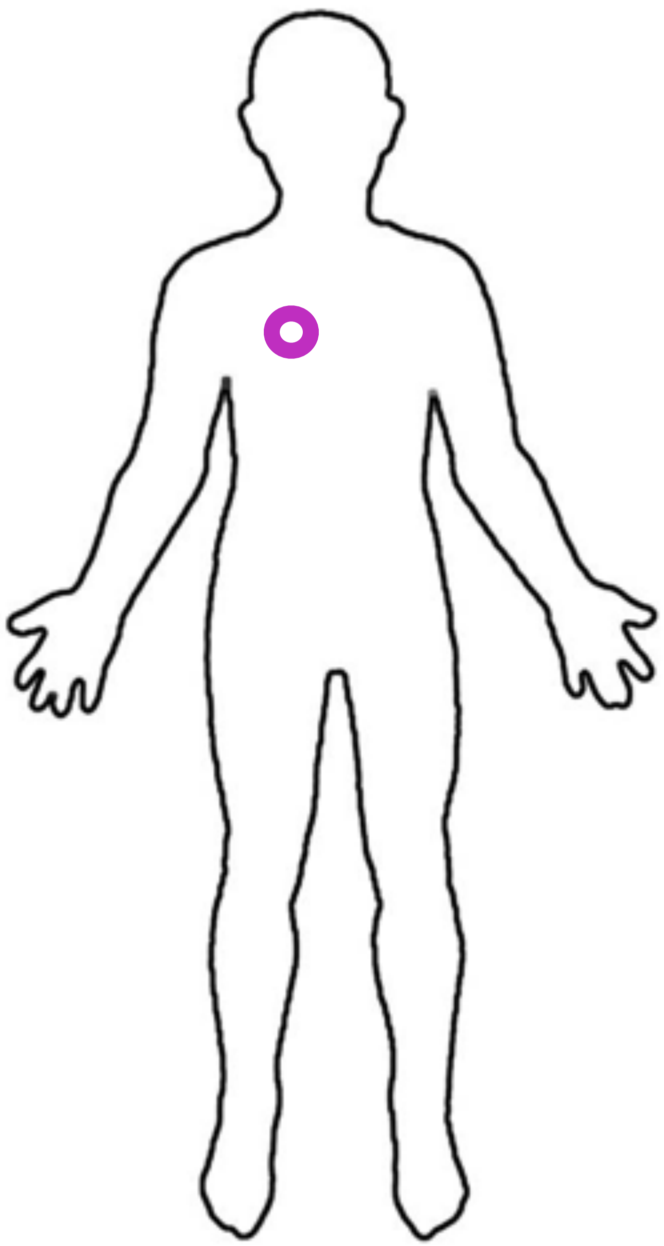
It causes cancer [Baylin & Ohm (2006)]

- Hypomethylation of oncogenes
- Hypermethylation of tumor suppressor genes

Cancer taxonomies

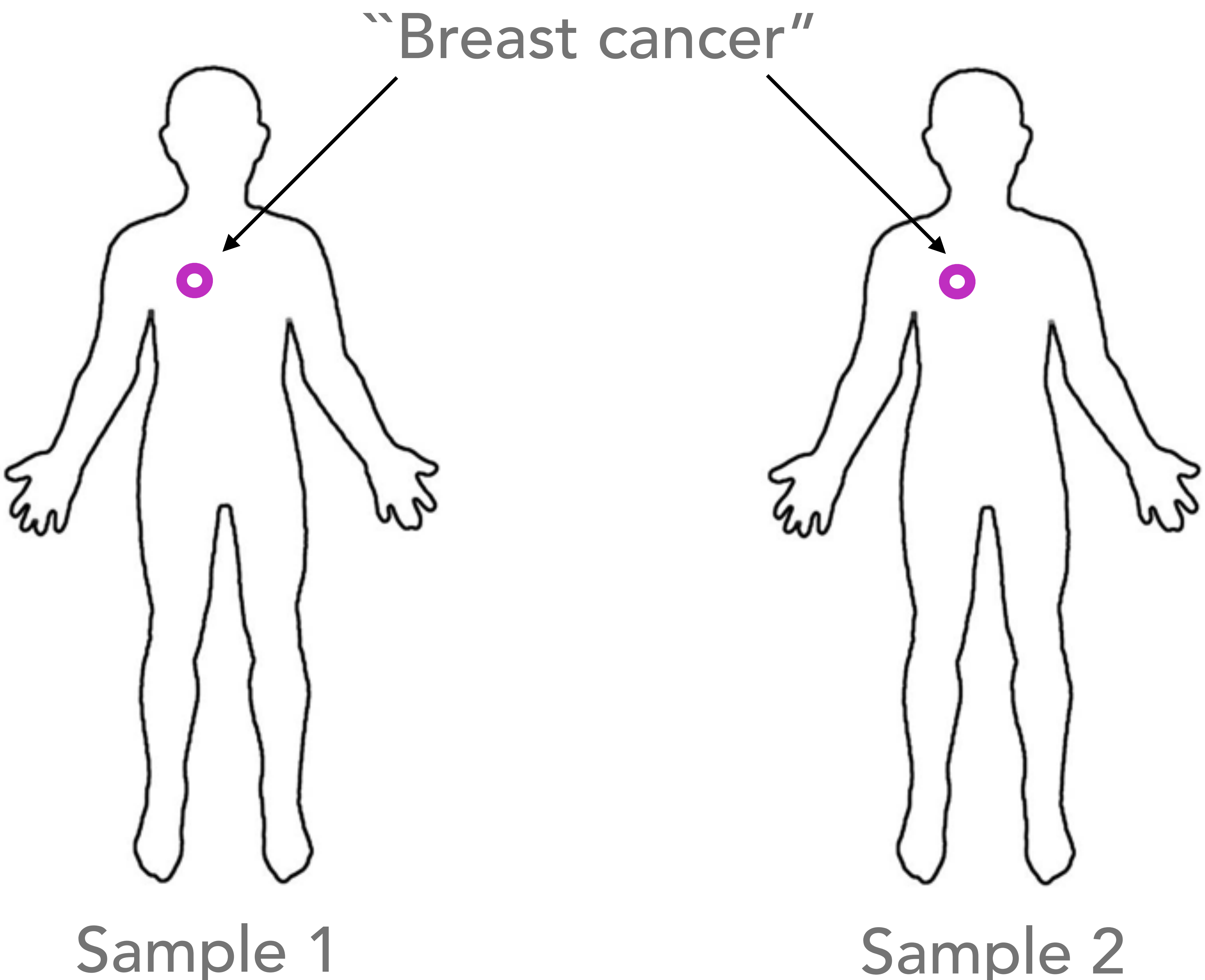


Sample 1

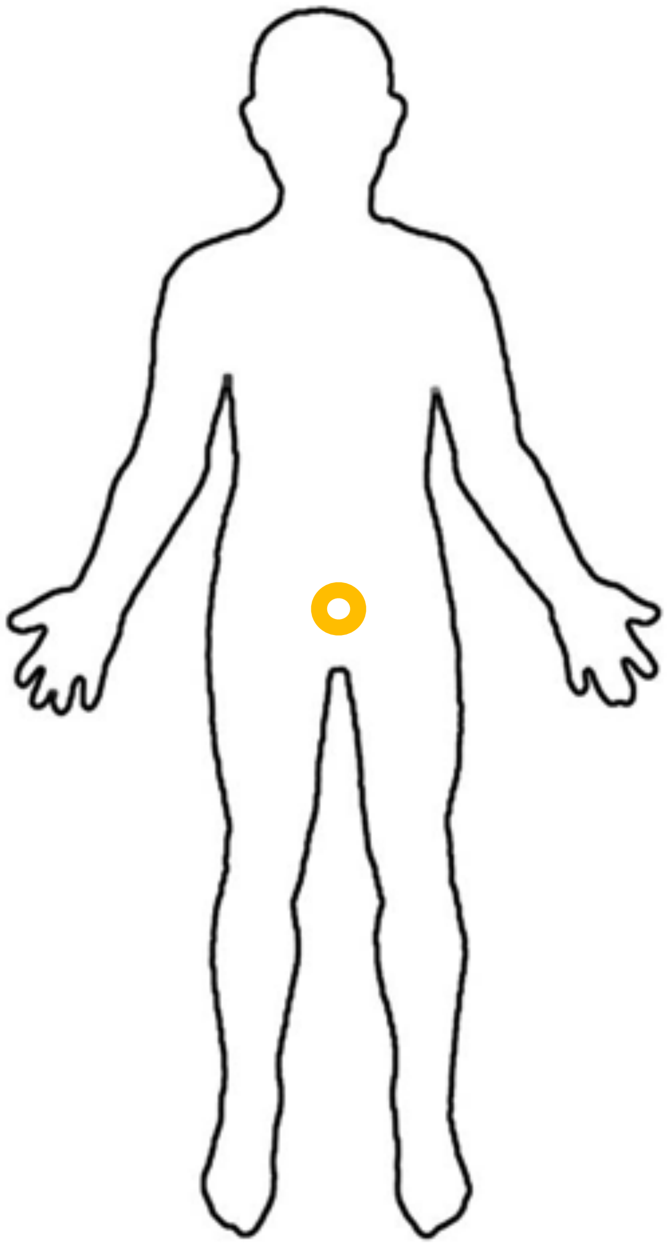


Sample 2

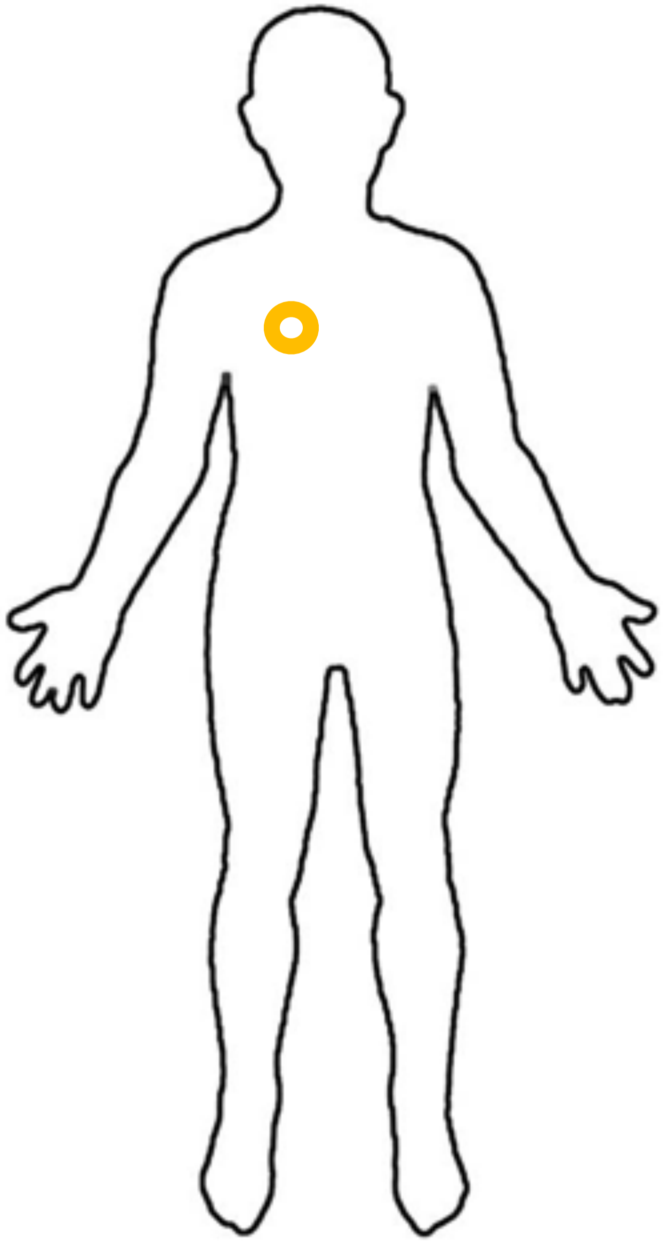
Cancer taxonomies



Cancer taxonomies

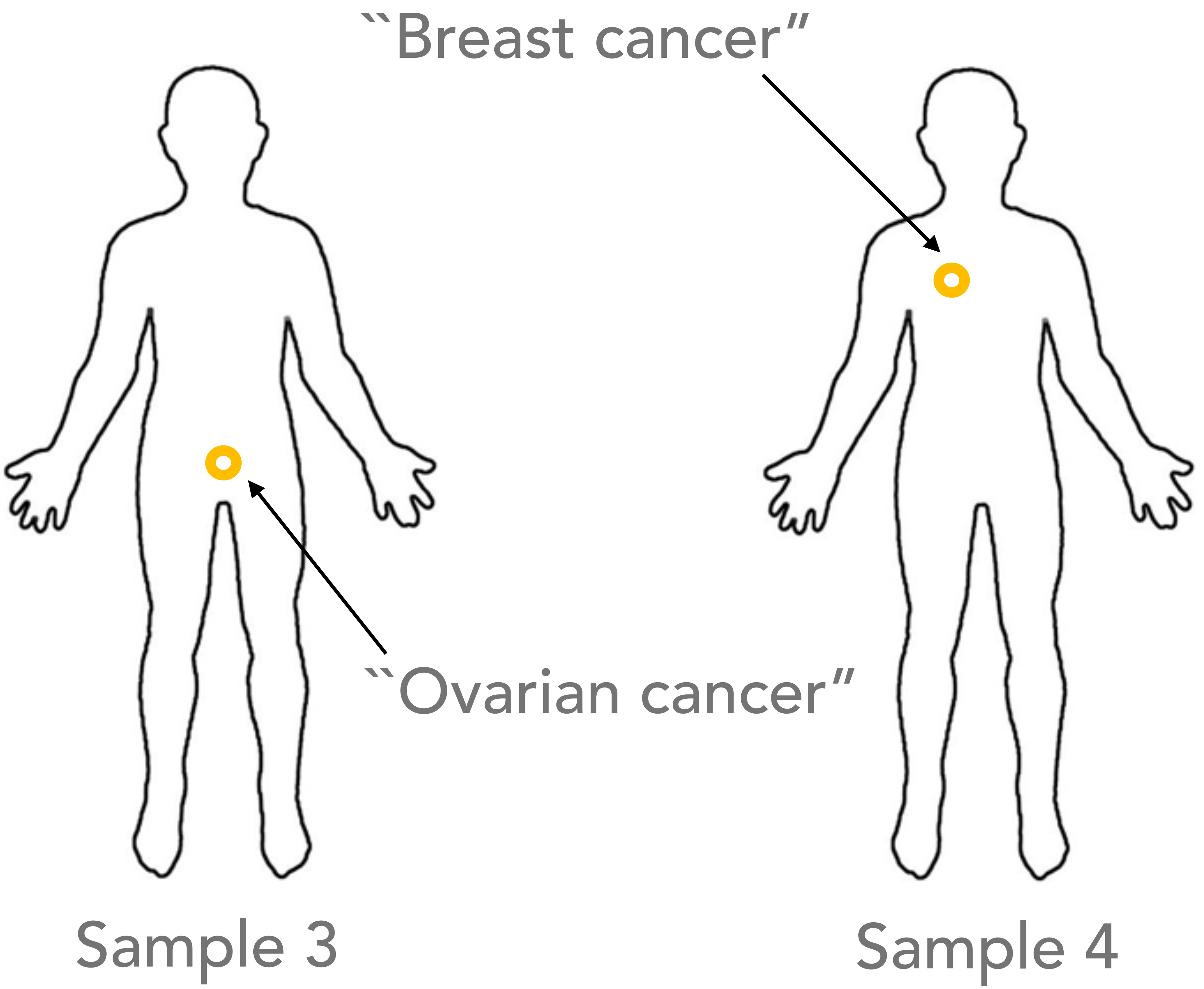


Sample 3



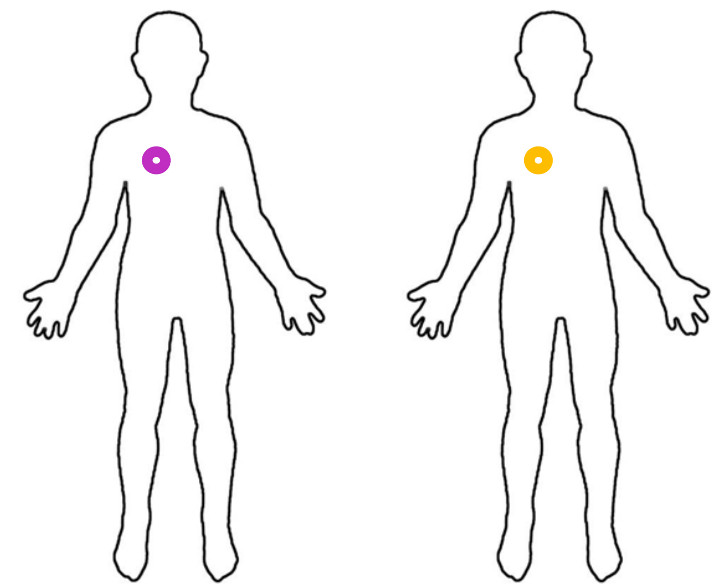
Sample 4

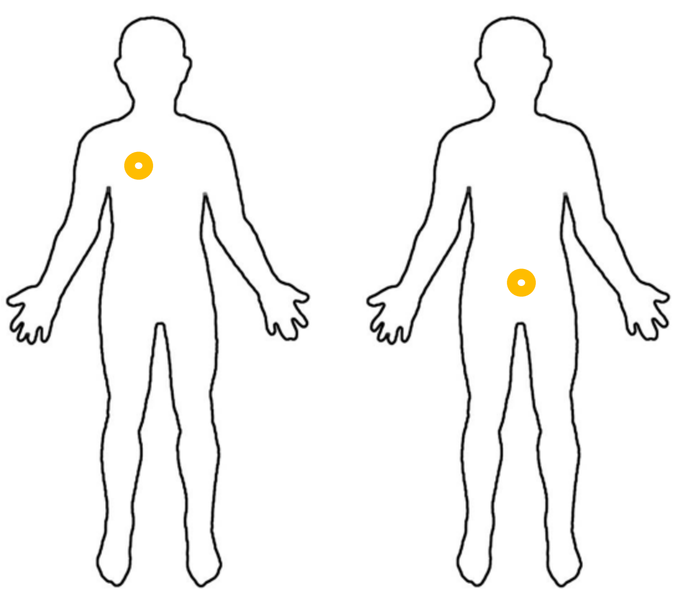
Cancer taxonomies



Cancer taxonomies

Anatomically similar cancer cells
may be genetically different





Anatomically different cancer cells
may be genetically similar

Cancer taxonomies

Goal: Develop new taxonomies based on genetic information

ML solution: Unsupervised dimensionality reduction

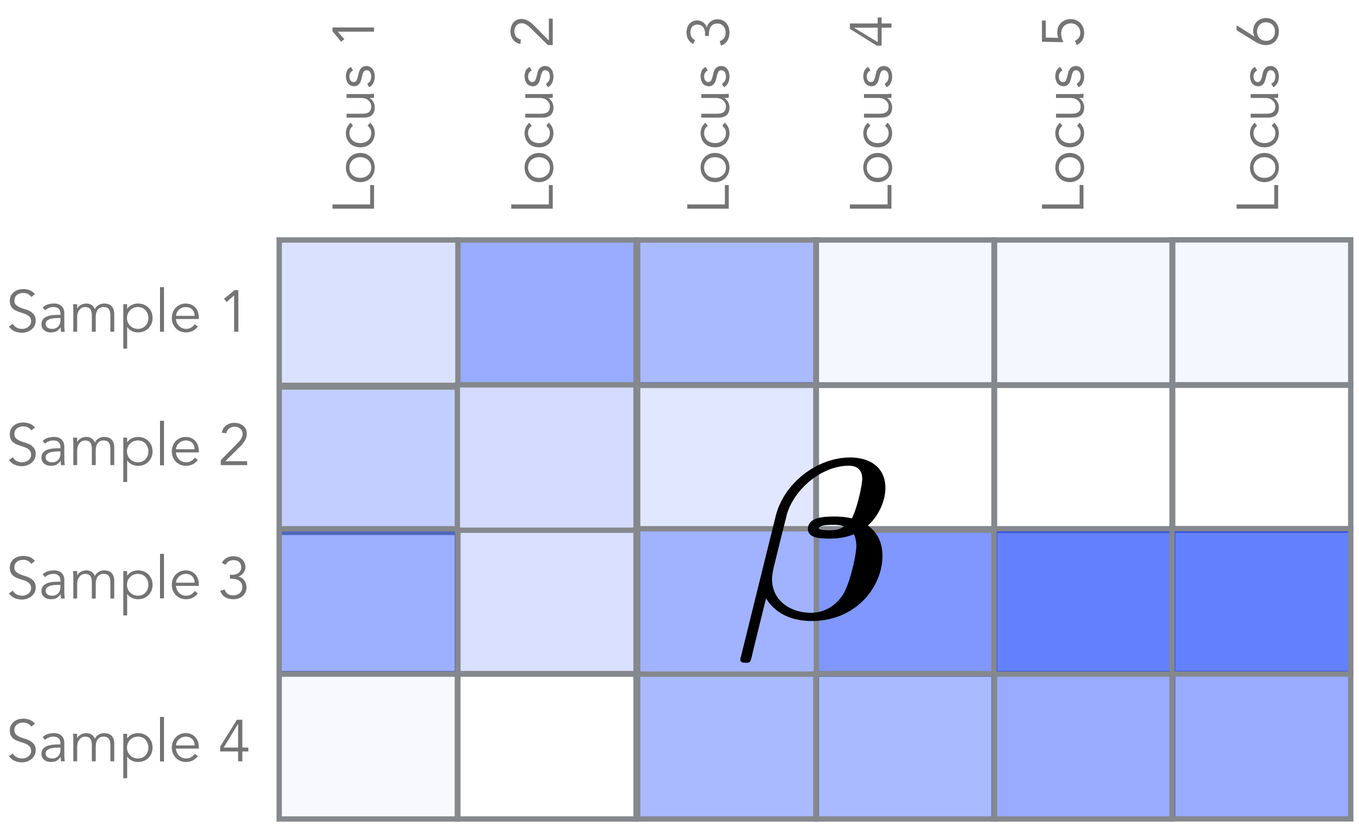
PCA, NMF, ICA,...

[Flusberg et al. (2010)]

[Wang et al. (2006)]

[Teschendorff et al. (2007)]

DNA methylation data

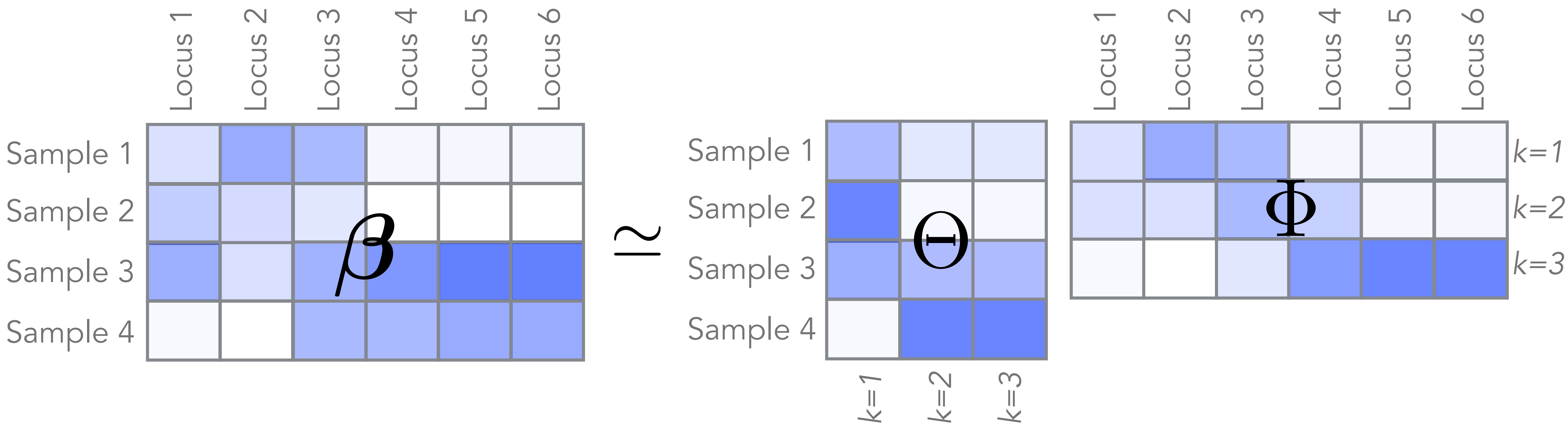


β_{ij} = how methylated locus *j* is in sample *i*

$$\beta_{ij} \in [0, 1]$$

CP decomposition

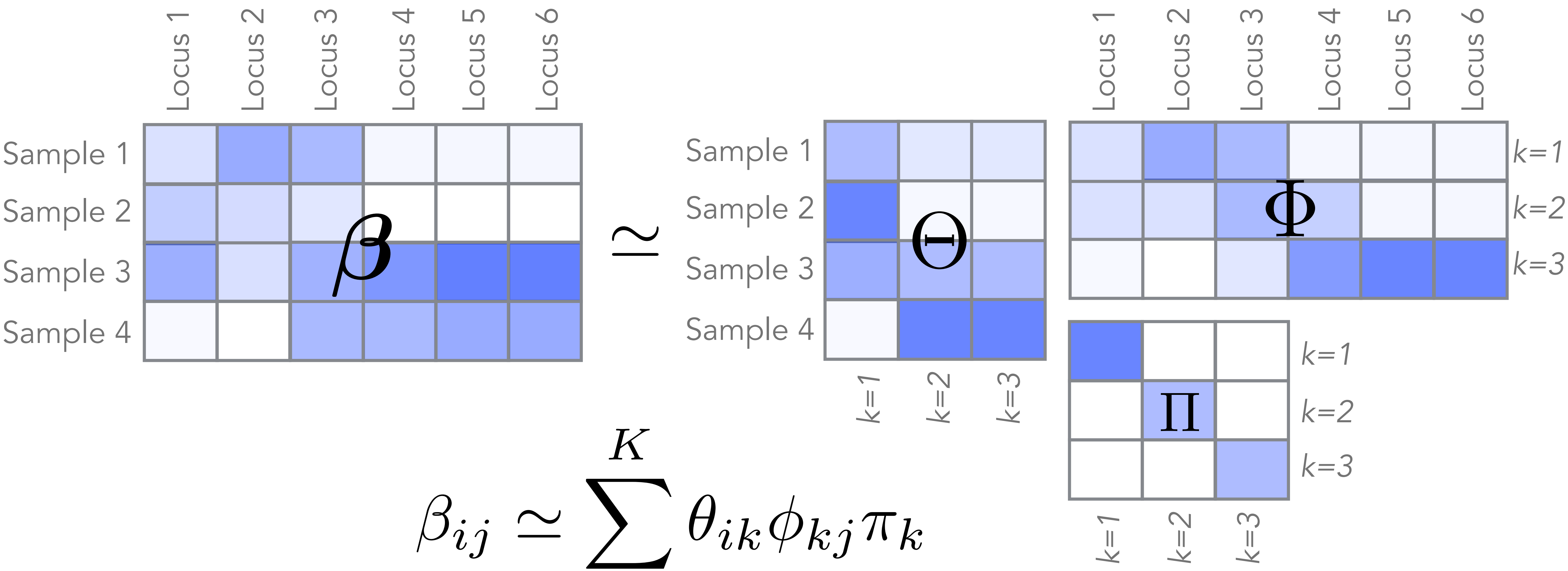
K "components"



$$\beta_{ij} \simeq \sum_{k=1}^{K} \theta_{ik}\phi_{kj}$$

CP decomposition

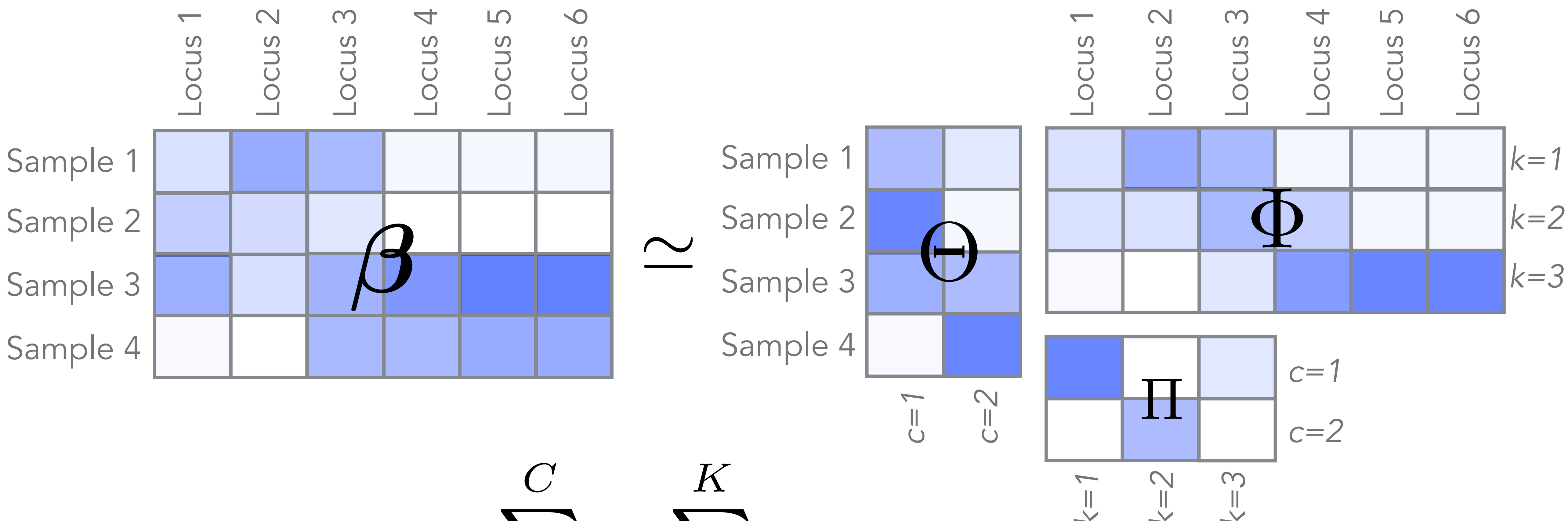
K “components”



$$\beta_{ij} \simeq \sum_{k=1}^{K} \theta_{ik}\phi_{kj}\pi_k$$

Tucker decomposition

C "clusters" and K "components"



$$\beta_{ij} \simeq \sum_{c=1}^{C} \theta_{ic} \sum_{k=1}^{K} \pi_{ck}\, \phi_{kj}$$

Beta Tucker decomposition

Our contributions:

- Novel generative model
  - Based on the Tucker decomposition
  - Matches the true data-generating process
    - ✓ Beta likelihood [Ma et al. (2015)]
    - ✓ Latent variables match real ones
    - ✓ Priors match known sources of noise
- Gibbs sampler with closed form conditionals

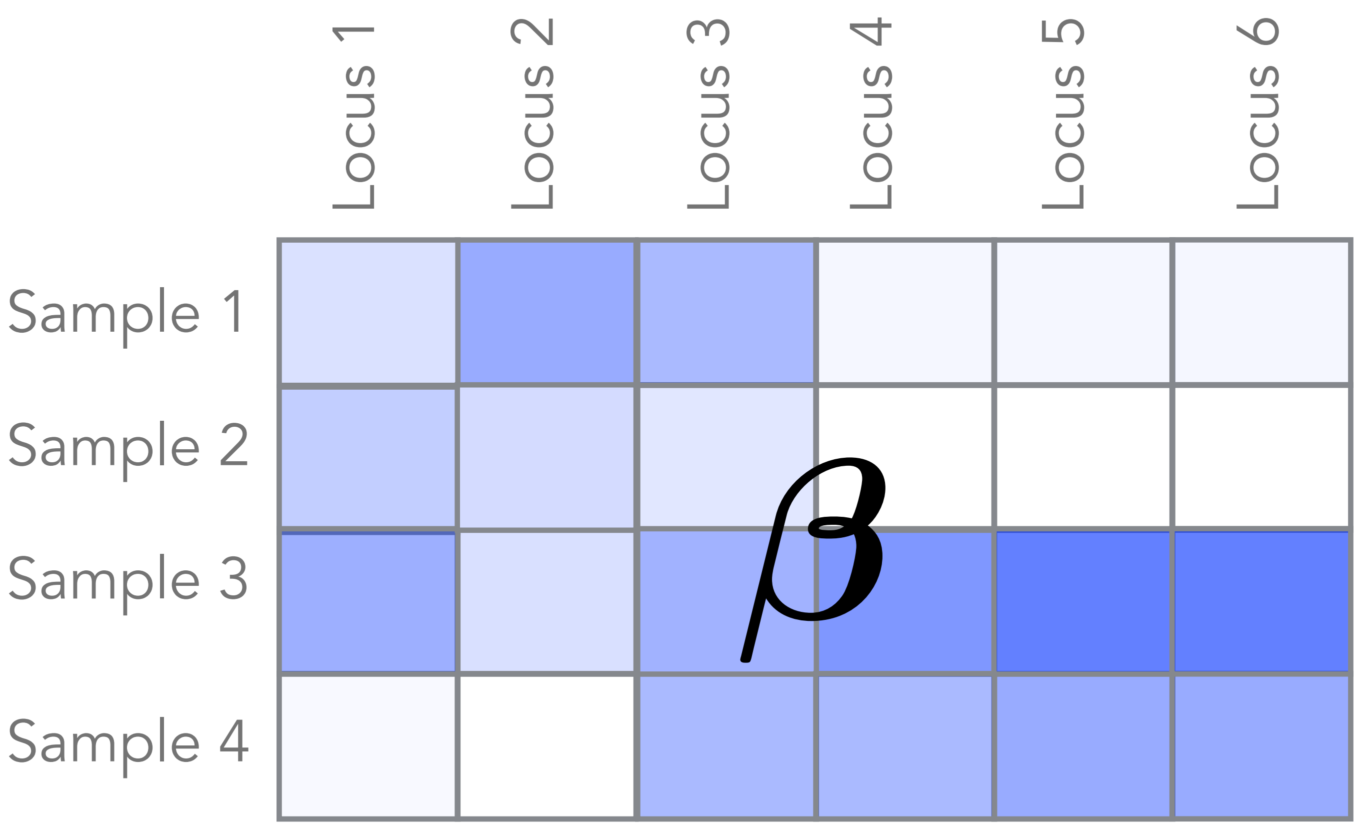
Is it better than PCA/NMF/ICA/etc in theory?

- Yes

Is it better than PCA/NMF/ICA/etc in practice?

- Comparable performance on (contrived) prediction tasks
- ??

DNA methylation data



β_{ij} = how methylated locus *j* is in sample *i*

$$\beta_{ij} \in [0, 1]$$

DNA methylation data

M M M M M M

CGCCTTCTCTCC**CG**A**CGCG**CCTTCTCT CTCC**CGCG**TCC**CGCG**A**CG**CCTTCTTTTT**CG**TTTT

Sample *i*

DNA methylation data

M
CGCCTTCTCTCCCG

M
CGCCTTCTTTTT

M M M
CTCCCGCGTCCCGCGA
Locus *j*

M
ACGCGCCTTCTCT

Sample *i*

CGTTTTTCTC

DNA methylation data

$y_{ij}^{(m)}$ = num. of methylated CpG sites in locus j of sample

$y_{ij}^{(u)}$ = num. of unmethylated CpG sites in locus j of sample

M
CGCCTTCTCTCCCG

M
CGCCTTCTTTTT

M M M
CTCCCGCGTCCCGCGA
Locus j

M
ACGCGCCTTCTCT

Sample i

CGTTTTTCTC

DNA methylation data

M
CGCCTTCTCTCC**CG**

M M M
CTCC**CGCG**TCC**CGCG**A
Locus *j*

M
CGCCTTCTTTTT

M
A**CGCG**CCTTCTCT

Sample *i*

CGTTTTTCTC

DNA methylation data

[Wang & Petronis (2008)]

M
CGCCTTCTCTCCCG

M M M
CTCCCGCGTCCCGCGA
Locus j

M
CGCCTTCTTTTT

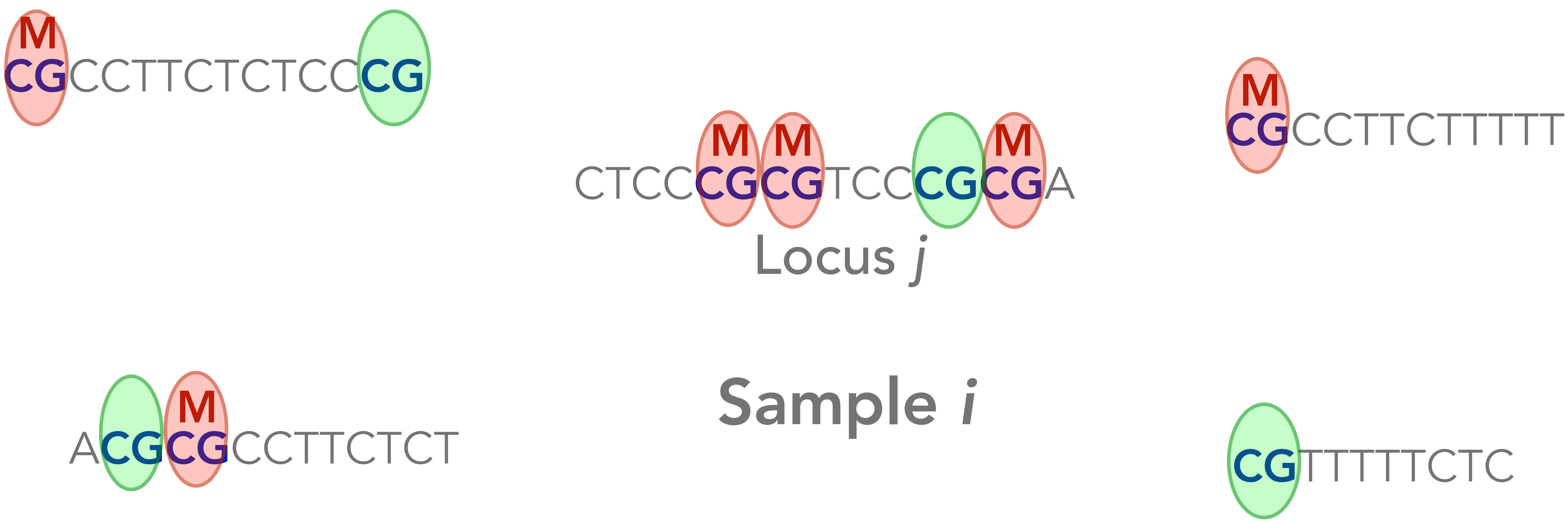
M
ACGCGCCTTCTCT

Sample i

CGTTTTTCTC

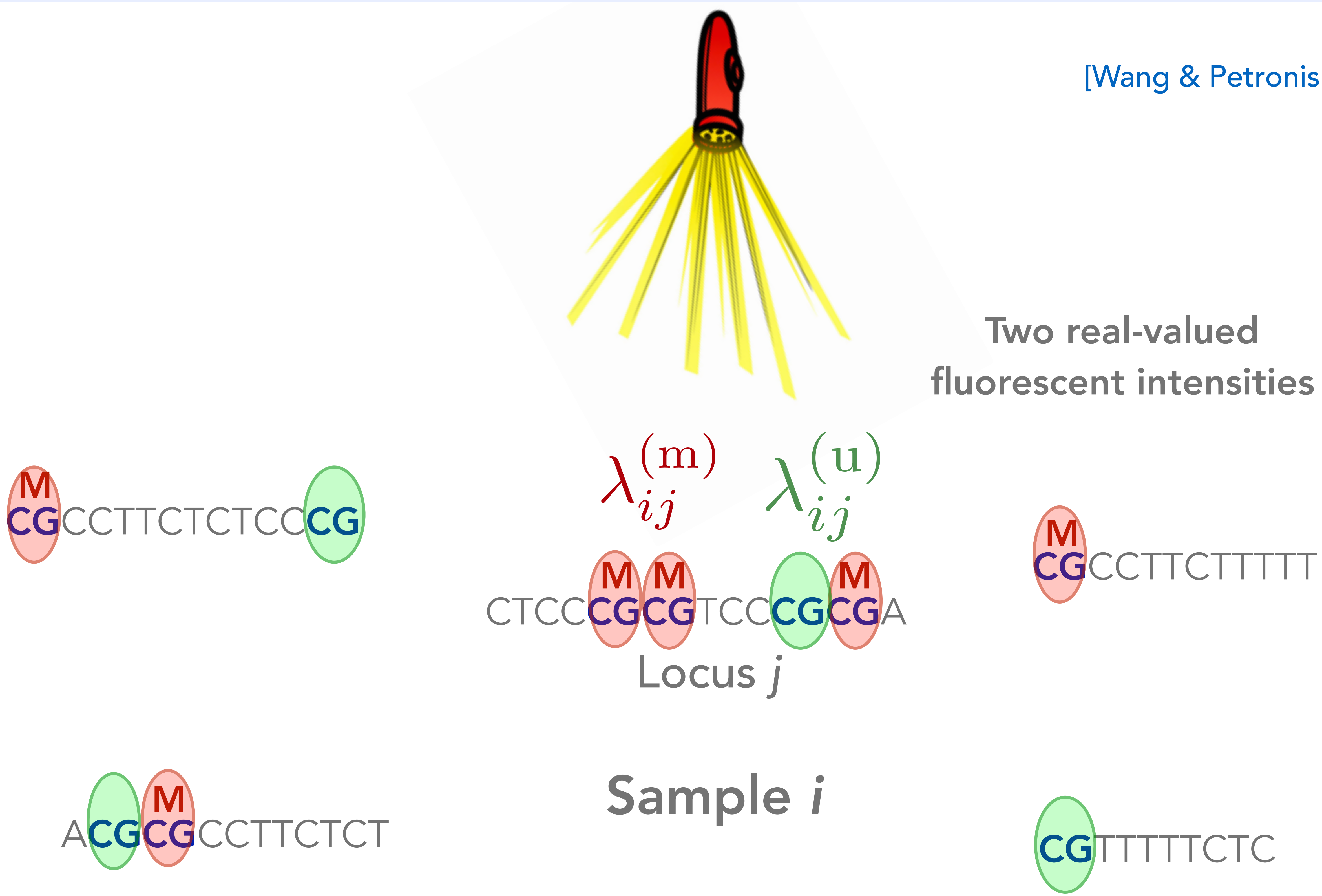
DNA methylation data

[Wang & Petronis (2008)]



DNA methylation data

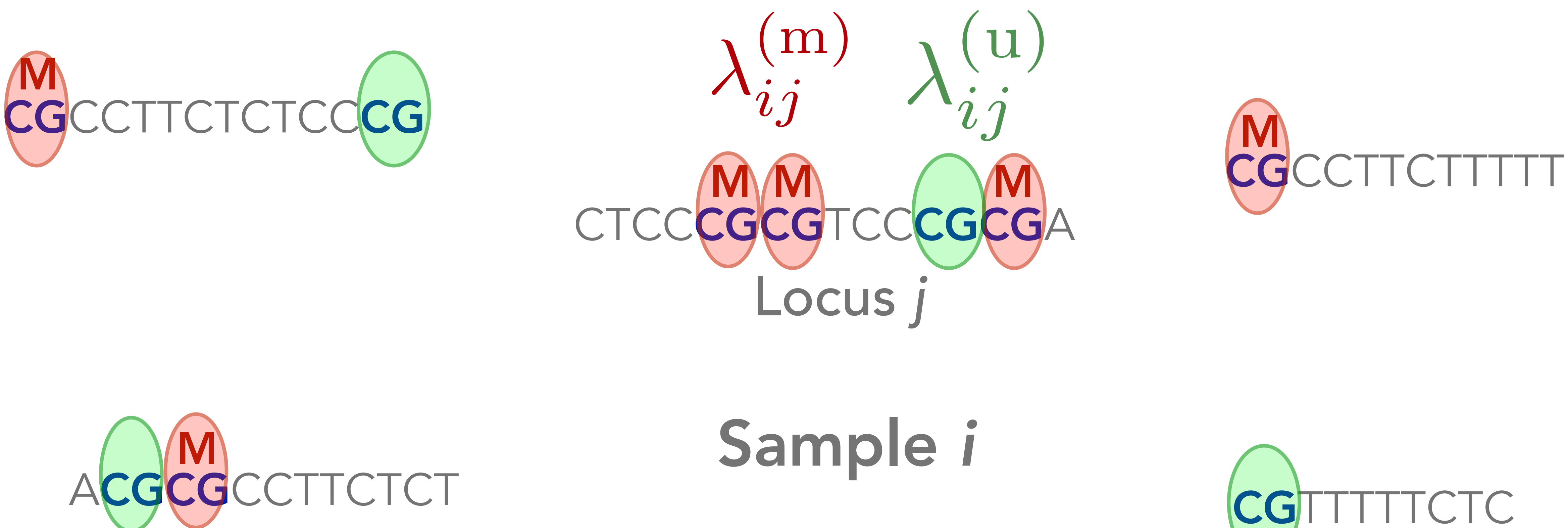
[Wang & Petronis (2008)]



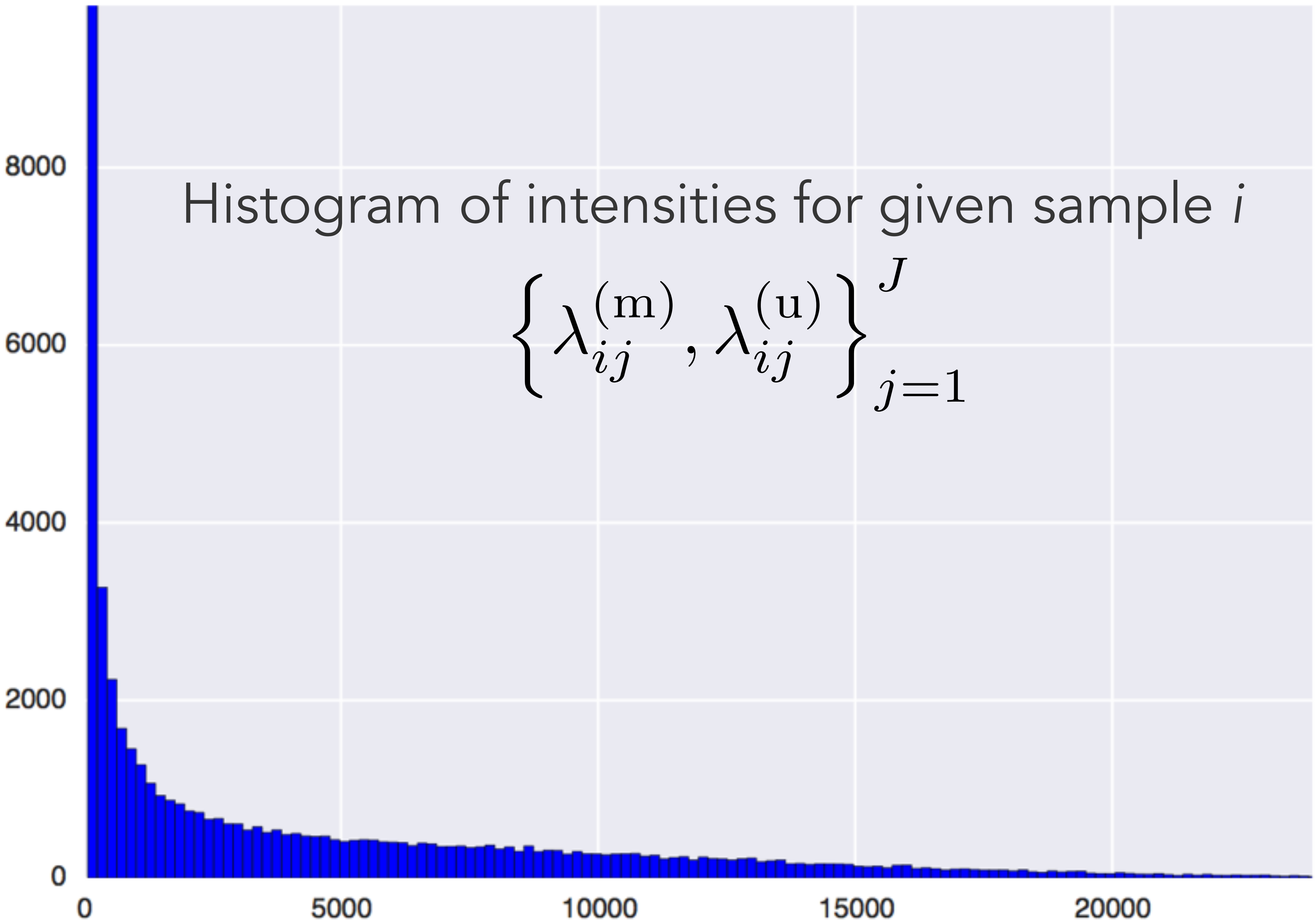
DNA methylation data

$$\beta_{ij} := \frac{\lambda_{ij}^{(\mathrm{m})}}{\lambda_{ij}^{(\mathrm{m})} + \lambda_{ij}^{(\mathrm{u})}}$$

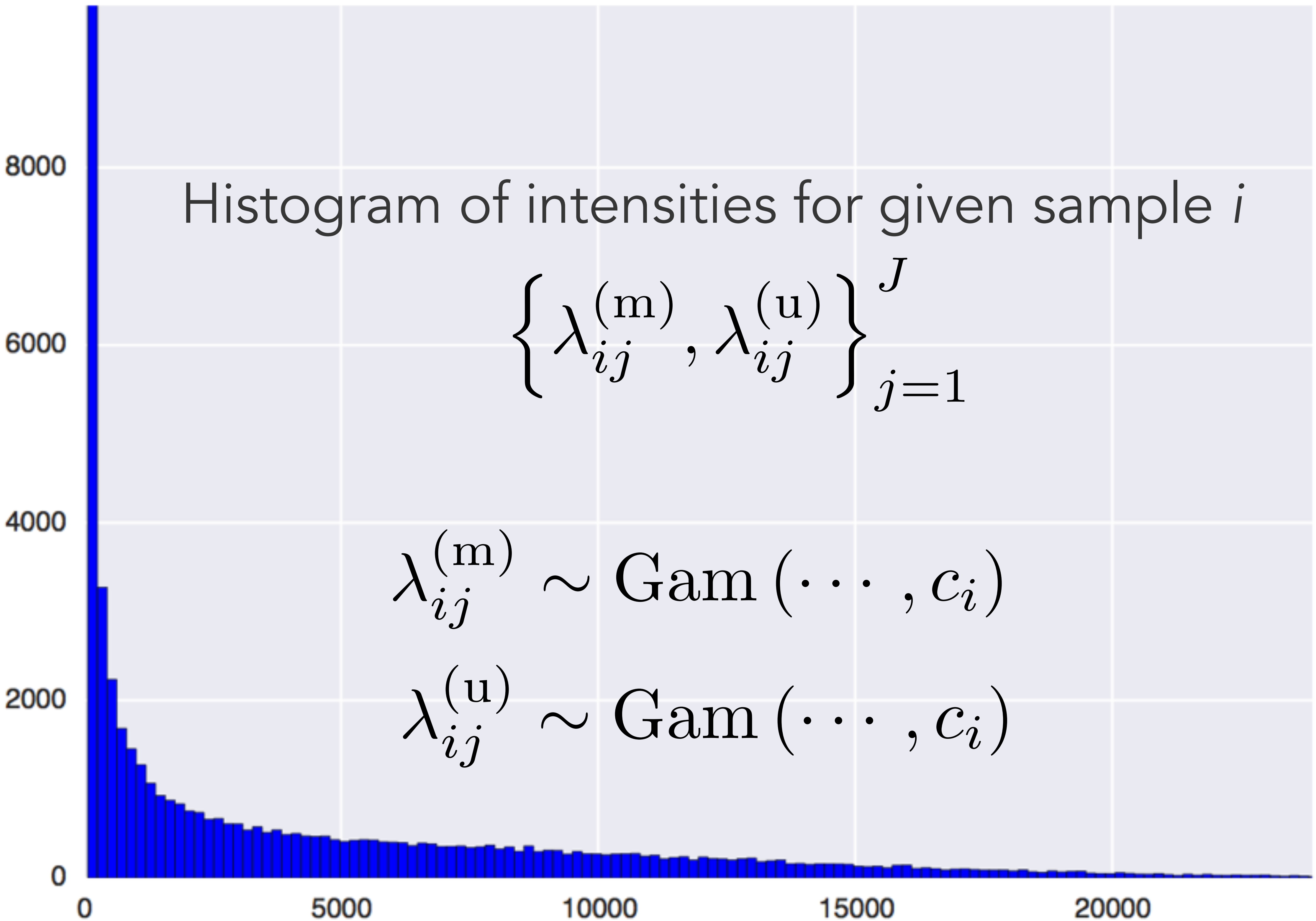
``Beta value"



DNA methylation data



DNA methylation data



Histogram of intensities for given sample *i*

$$\left\{ \lambda_{ij}^{(\mathrm{m})}, \lambda_{ij}^{(\mathrm{u})} \right\}_{j=1}^{J}$$

$$\lambda_{ij}^{(\mathrm{m})} \sim \mathrm{Gam}\left(\cdots, c_i\right)$$

$$\lambda_{ij}^{(\mathrm{u})} \sim \mathrm{Gam}\left(\cdots, c_i\right)$$

Gamma-Beta relationship

$$\lambda_1 \sim \mathrm{Gam}(\alpha_1,\, c) \qquad \lambda_2 \sim \mathrm{Gam}(\alpha_2,\, c)$$

$$\left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right) \sim \mathrm{Beta}(\alpha_1,\, \alpha_2)$$

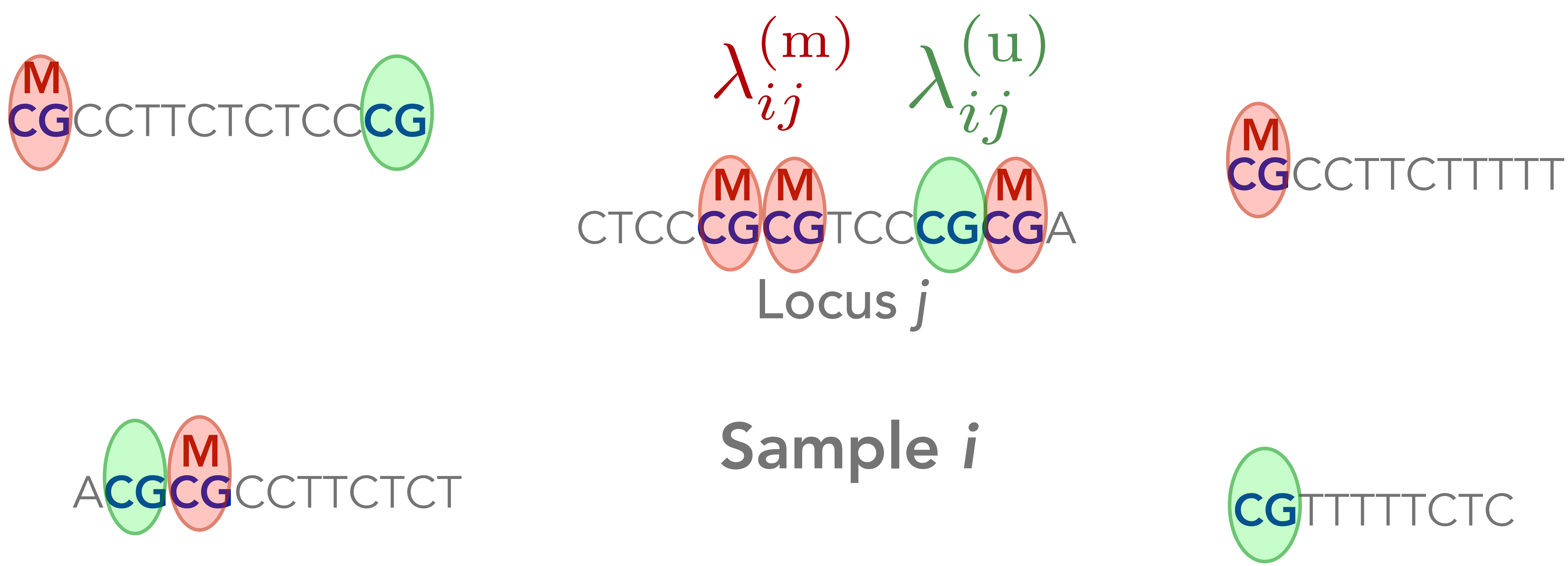
Gamma-Beta relationship

$$\lambda_{ij}^{(\mathrm{m})} \sim \mathrm{Gam}\left(\cdots, c_i\right) \qquad \lambda_{ij}^{(\mathrm{u})} \sim \mathrm{Gam}\left(\cdots, c_i\right)$$

$$\beta_{ij} := \frac{\lambda_{ij}^{(\mathrm{m})}}{\lambda_{ij}^{(\mathrm{m})} + \lambda_{ij}^{(\mathrm{u})}}$$

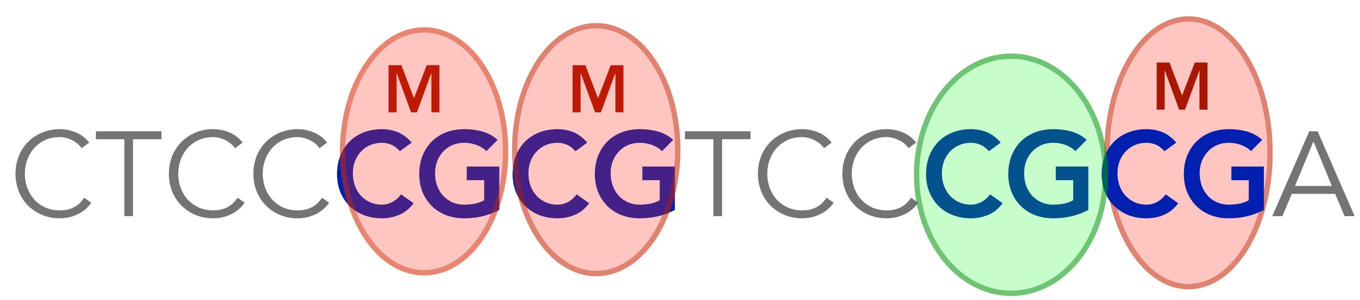
$$\beta_{ij} \sim \mathrm{Beta}(\cdots, \cdots)$$

Beta Tucker decomposition



Beta Tucker decomposition

$\lambda_{ij}^{(\mathrm{m})}$ $\lambda_{ij}^{(\mathrm{u})}$

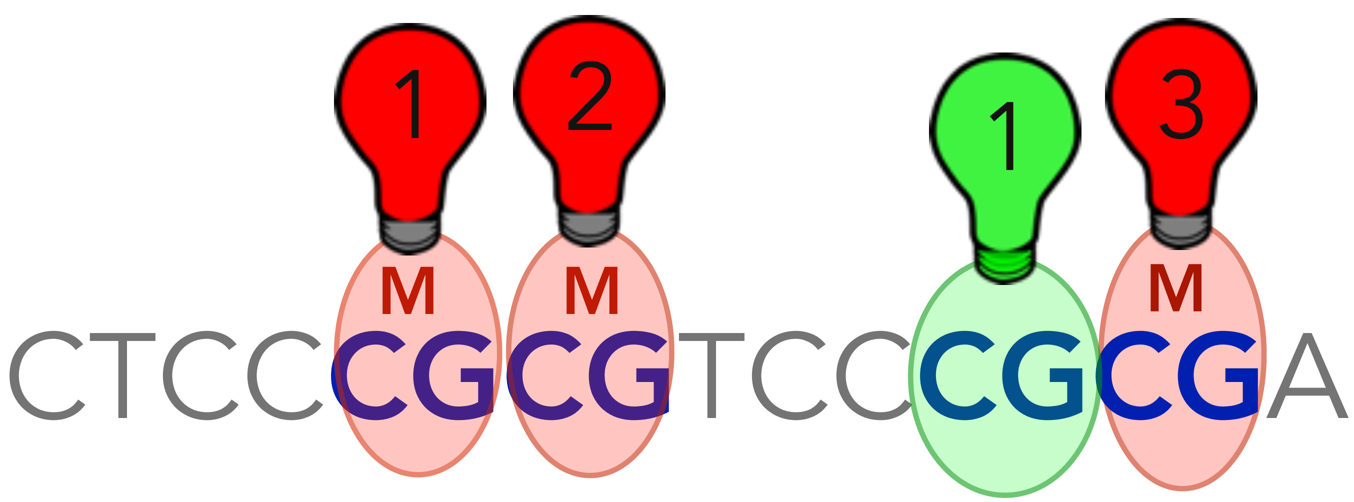


Locus j

Sample i

Beta Tucker decomposition

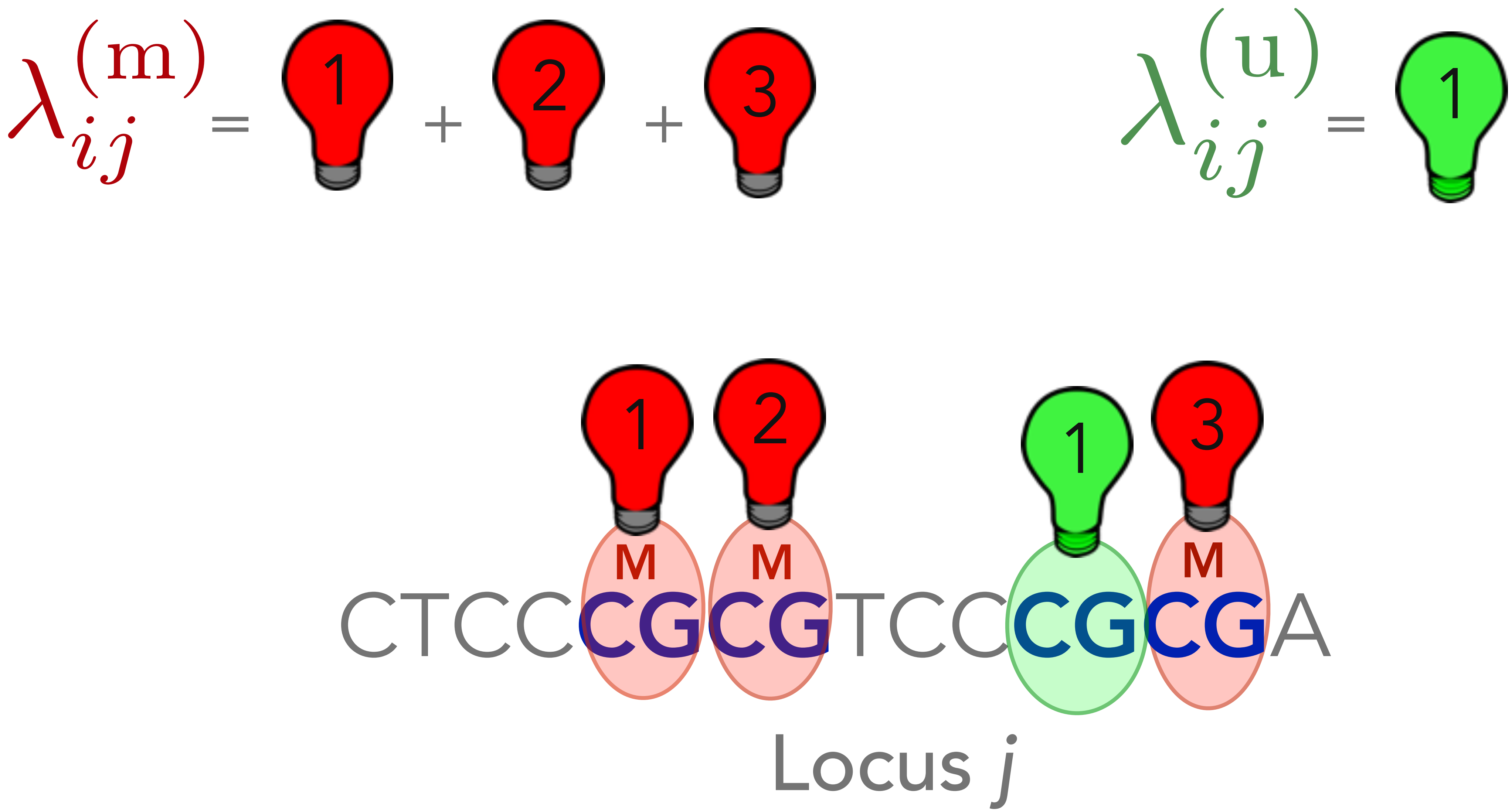
$\lambda_{ij}^{(\mathrm{m})}$ $\lambda_{ij}^{(\mathrm{u})}$



Locus j

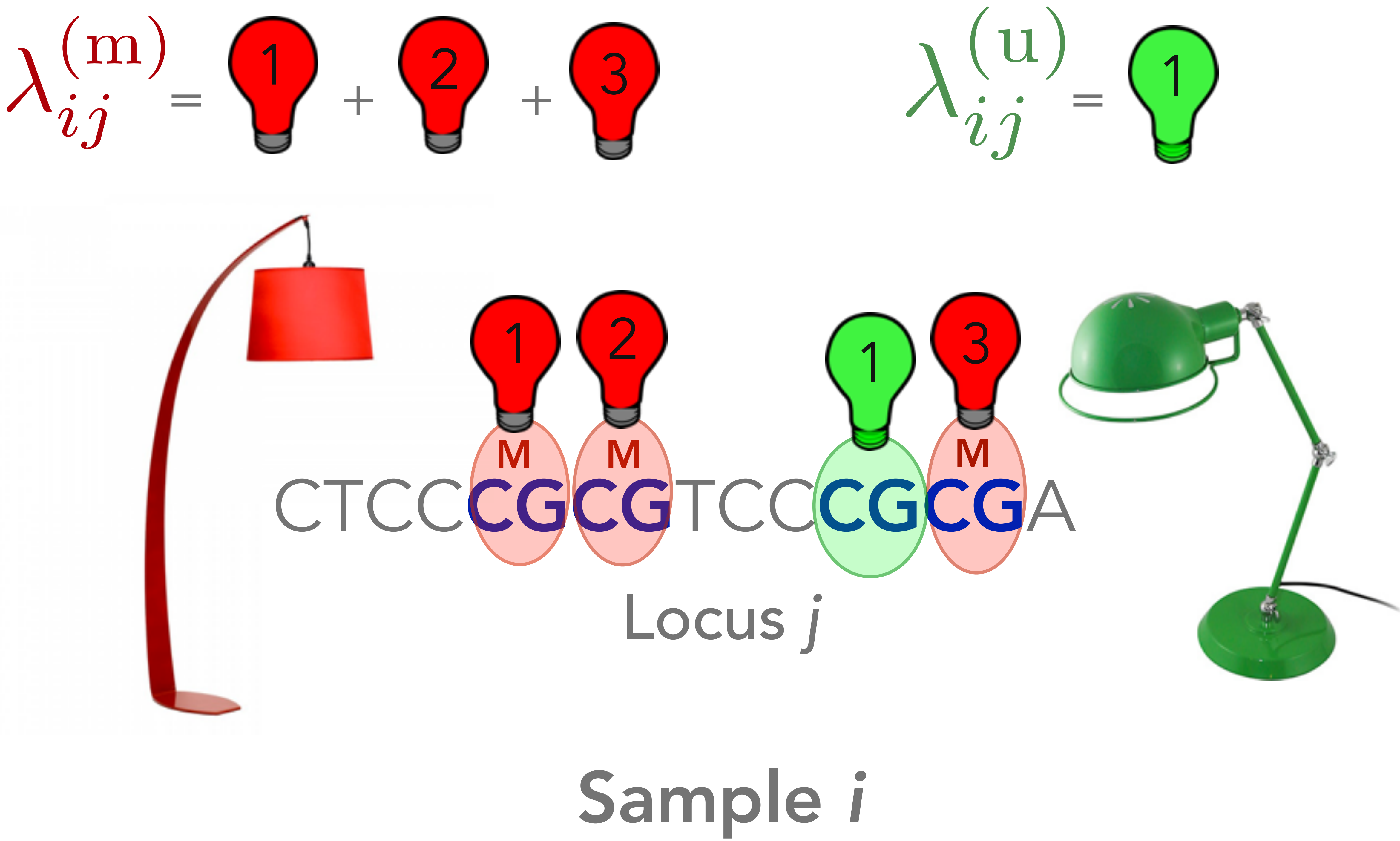
Sample $\boldsymbol{i}$

Beta Tucker decomposition

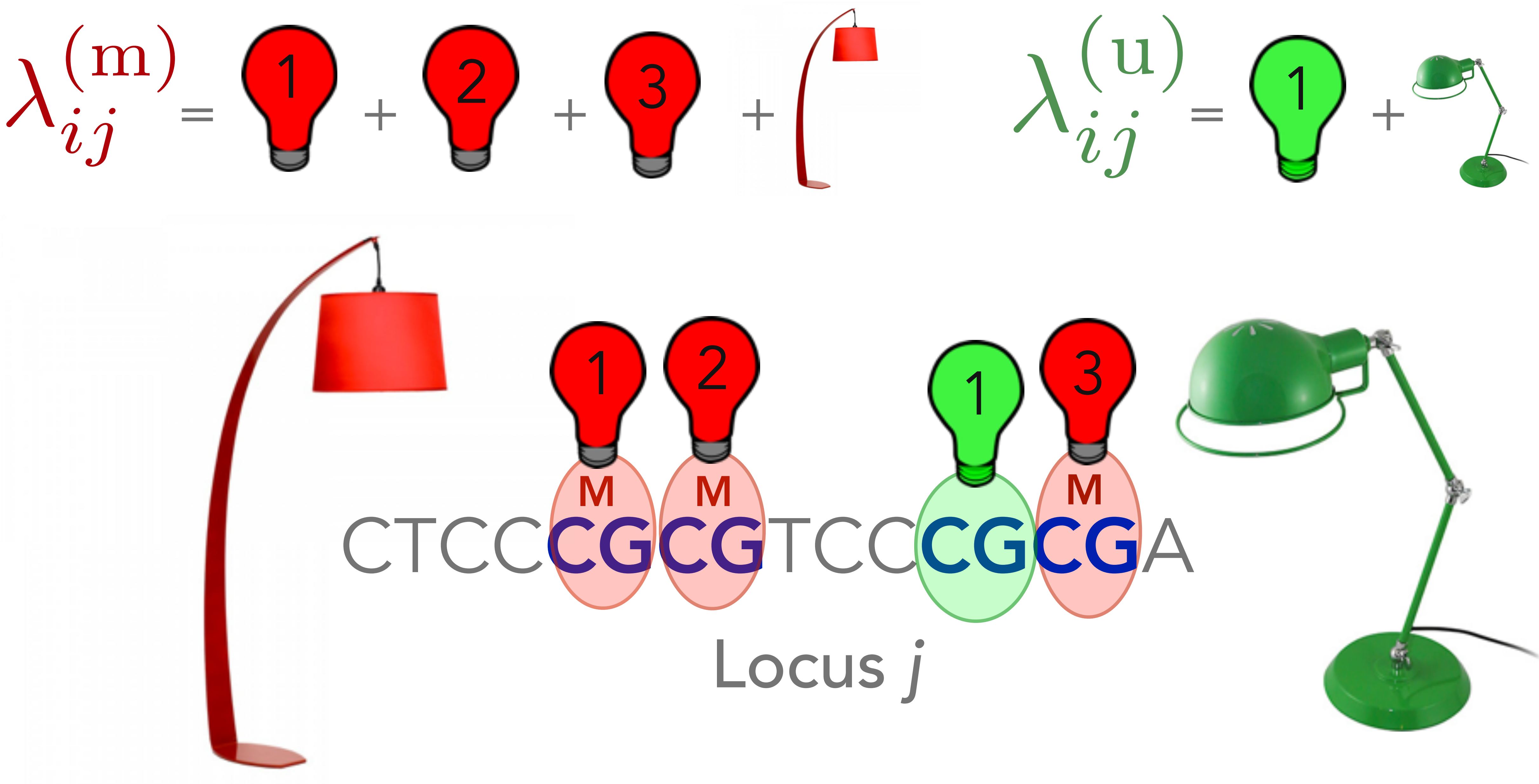


Sample i

Beta Tucker decomposition



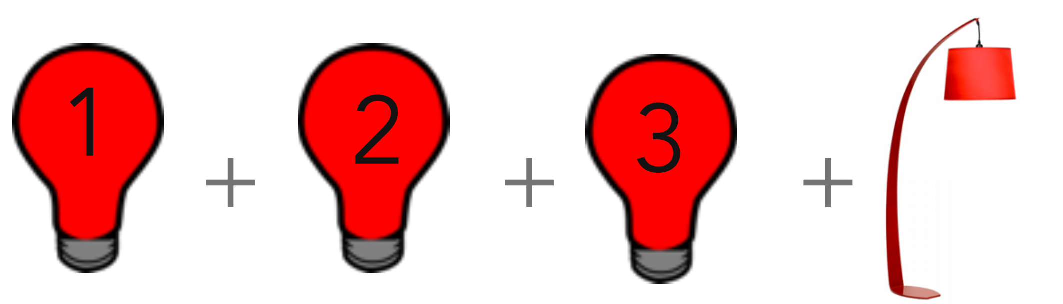
Beta Tucker decomposition



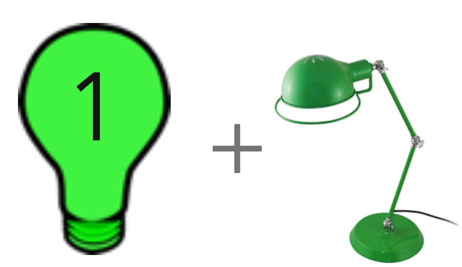
Sample *i*

Beta Tucker decomposition

$$\lambda_{ij}^{(\mathrm{m})} =$$



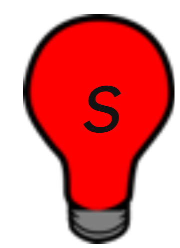
$$\lambda_{ij}^{(\mathrm{u})} =$$



Locus *j*

Sample *i*

Beta Tucker decomposition

$\lambda_{ij}^{(m)} =$  $+ \sum_{s=1}^{y_{ij}^{(m)}}$ 

$\lambda_{ij}^{(u)} =$  $+ \sum_{s=1}^{y_{ij}^{(u)}}$ 

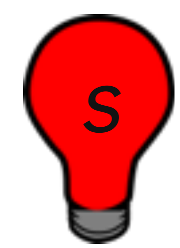
Locus *j*

Sample *i*

Beta Tucker decomposition

$\sim \mathrm{Gam}(b_0, c_i)$

$\sim \mathrm{Gam}(1, c_i)$

$\lambda_{ij}^{(\mathrm{m})} =$  $+ \sum_{s=1}^{y_{ij}^{(\mathrm{m})}}$ 

$\lambda_{ij}^{(\mathrm{u})} =$  $+ \sum_{s=1}^{y_{ij}^{(\mathrm{u})}}$ 

Locus *j*

Sample *i*

Beta Tucker decomposition

$$\lambda_{ij}^{(\mathrm{m})} \sim \mathrm{Gam}\left(b_0 + y_{ij}^{(\mathrm{m})},\ c_i\right)$$

$$\lambda_{ij}^{(\mathrm{u})} \sim \mathrm{Gam}\left(b_0 + y_{ij}^{(\mathrm{u})},\ c_i\right)$$

Locus *j*

Sample *i*

Beta Tucker decomposition

$$\lambda_{ij}^{(\mathrm{m})} \sim \mathrm{Gam}\left(b_0 + y_{ij}^{(\mathrm{m})},\, c_i\right) \qquad \lambda_{ij}^{(\mathrm{u})} \sim \mathrm{Gam}\left(b_0 + y_{ij}^{(\mathrm{u})},\, c_i\right)$$

$$\beta_{ij} := \frac{\lambda_{ij}^{(\mathrm{m})}}{\lambda_{ij}^{(\mathrm{m})} + \lambda_{ij}^{(\mathrm{u})}}$$

Equivalent to:

$$\beta_{ij} \sim \mathrm{Beta}\left(b_0 + y_{ij}^{(\mathrm{m})},\, b_0 + y_{ij}^{(\mathrm{u})}\right)$$

Beta Tucker decomposition

$$y_{ij}^{(\mathrm{m})} \sim \mathrm{Pois}(\cdots) \qquad y_{ij}^{(\mathrm{u})} \sim \mathrm{Pois}(\cdots)$$

$$\lambda_{ij}^{(\mathrm{m})} \sim \mathrm{Gam}\left(b_0 + y_{ij}^{(\mathrm{m})},\, c_i\right) \qquad \lambda_{ij}^{(\mathrm{u})} \sim \mathrm{Gam}\left(b_0 + y_{ij}^{(\mathrm{u})},\, c_i\right)$$

$$\beta_{ij} := \frac{\lambda_{ij}^{(\mathrm{m})}}{\lambda_{ij}^{(\mathrm{m})} + \lambda_{ij}^{(\mathrm{u})}}$$

Beta Tucker decomposition

$$y_{ij}^{(\mathrm{m})} \sim \mathrm{Pois}\left(\gamma \sum_{c=1}^{C} \theta_{ic} \sum_{k=1}^{K} \pi_{ck}\, \phi_{kj}\right)$$

Beta Tucker decomposition

$$y_{ij}^{(\mathrm{m})} \sim \mathrm{Pois}\left(\gamma \sum_{c=1}^{C} \theta_{ic} \sum_{k=1}^{K} \pi_{ck}\, \phi_{kj}\right)$$

the **probability** that
sample *i* is in cluster *c*

Beta Tucker decomposition

$$y_{ij}^{(\mathrm{m})} \sim \mathrm{Pois}\left(\gamma \sum_{c=1}^{C} \theta_{ic} \sum_{k=1}^{K} \pi_{ck}\, \phi_{kj}\right)$$

the **probability** that
samples in cluster *c*
methylate loci in component *k*

Beta Tucker decomposition

$$y_{ij}^{(\mathrm{m})} \sim \mathrm{Pois}\left(\gamma \sum_{c=1}^{C} \theta_{ic} \sum_{k=1}^{K} \pi_{ck}\, \phi_{kj}\right)$$

the **probability** that
locus *j* is in component *k*

Beta Tucker decomposition

$$y_{ij}^{(\mathrm{m})} \sim \mathrm{Pois}\left(\gamma \sum_{c=1}^{C} \theta_{ic} \sum_{k=1}^{K} \pi_{ck}\, \phi_{kj}\right)$$

$$\boldsymbol{\theta}_i \sim \mathrm{Dir}(\eta_1, \ldots, \eta_C)$$

$$\pi_{ck} \sim \mathrm{Beta}(\eta_0^{(\mathrm{m})},\, \eta_0^{(\mathrm{u})})$$

$$\boldsymbol{\phi}_j \sim \mathrm{Dir}(\nu_1, \ldots, \nu_K)$$

Beta Tucker decomposition

$$y_{ij}^{(\mathrm{m})} \sim \mathrm{Pois}\left(\gamma \underbrace{\sum_{c=1}^{C} \theta_{ic} \sum_{k=1}^{K} \pi_{ck}\, \phi_{kj}}_{=\, p_{ij}}\right)$$

Beta Tucker decomposition

$$y_{ij}^{(\mathrm{m})} \sim \mathrm{Pois}\left(\gamma \underbrace{\sum_{c=1}^{C} \theta_{ic} \sum_{k=1}^{K} \pi_{ck}\, \phi_{kj}}_{=\, p_{ij}}\right)$$

Beta Tucker decomposition

$$y_{ij}^{(\mathrm{m})} \sim \mathrm{Pois}(\gamma\, p_{ij})$$

the **probability** that
sample *i* methylates
CpG sites in locus *j*

Beta Tucker decomposition

$$y_{ij}^{(\mathrm{m})} \sim \mathrm{Pois}(\gamma\, p_{ij})$$

the occurrence rate
of CpG sites

Beta Tucker decomposition

$$p_{ij} := \sum_{c=1}^{C} \theta_{ic} \sum_{k=1}^{K} \pi_{ck} \phi_{kj}$$

$$y_{ij}^{(\mathrm{m})} \sim \mathrm{Pois}\big(\gamma\, p_{ij}\big) \qquad y_{ij}^{(\mathrm{u})} \sim \mathrm{Pois}\big(\gamma\, (1 - p_{ij})\big)$$

$$\lambda_{ij}^{(\mathrm{m})} \sim \mathrm{Gam}\left(b_0 + y_{ij}^{(\mathrm{m})},\, c_i\right) \qquad \lambda_{ij}^{(\mathrm{u})} \sim \mathrm{Gam}\left(b_0 + y_{ij}^{(\mathrm{u})},\, c_i\right)$$

$$\beta_{ij} := \frac{\lambda_{ij}^{(\mathrm{m})}}{\lambda_{ij}^{(\mathrm{m})} + \lambda_{ij}^{(\mathrm{u})}}$$

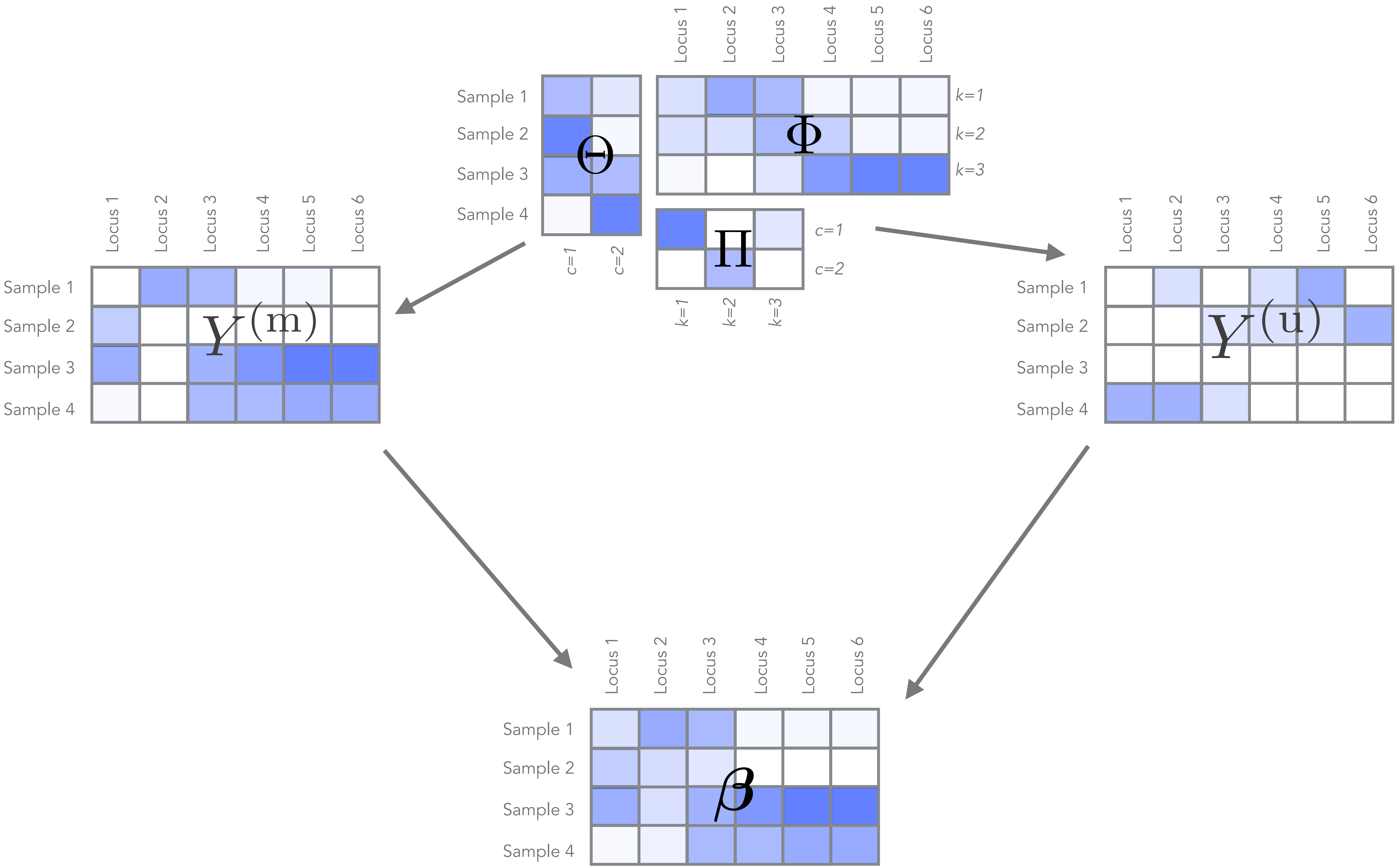
Beta Tucker decomposition

$$p_{ij} := \sum_{c=1}^{C} \theta_{ic} \sum_{k=1}^{K} \pi_{ck} \phi_{kj}$$

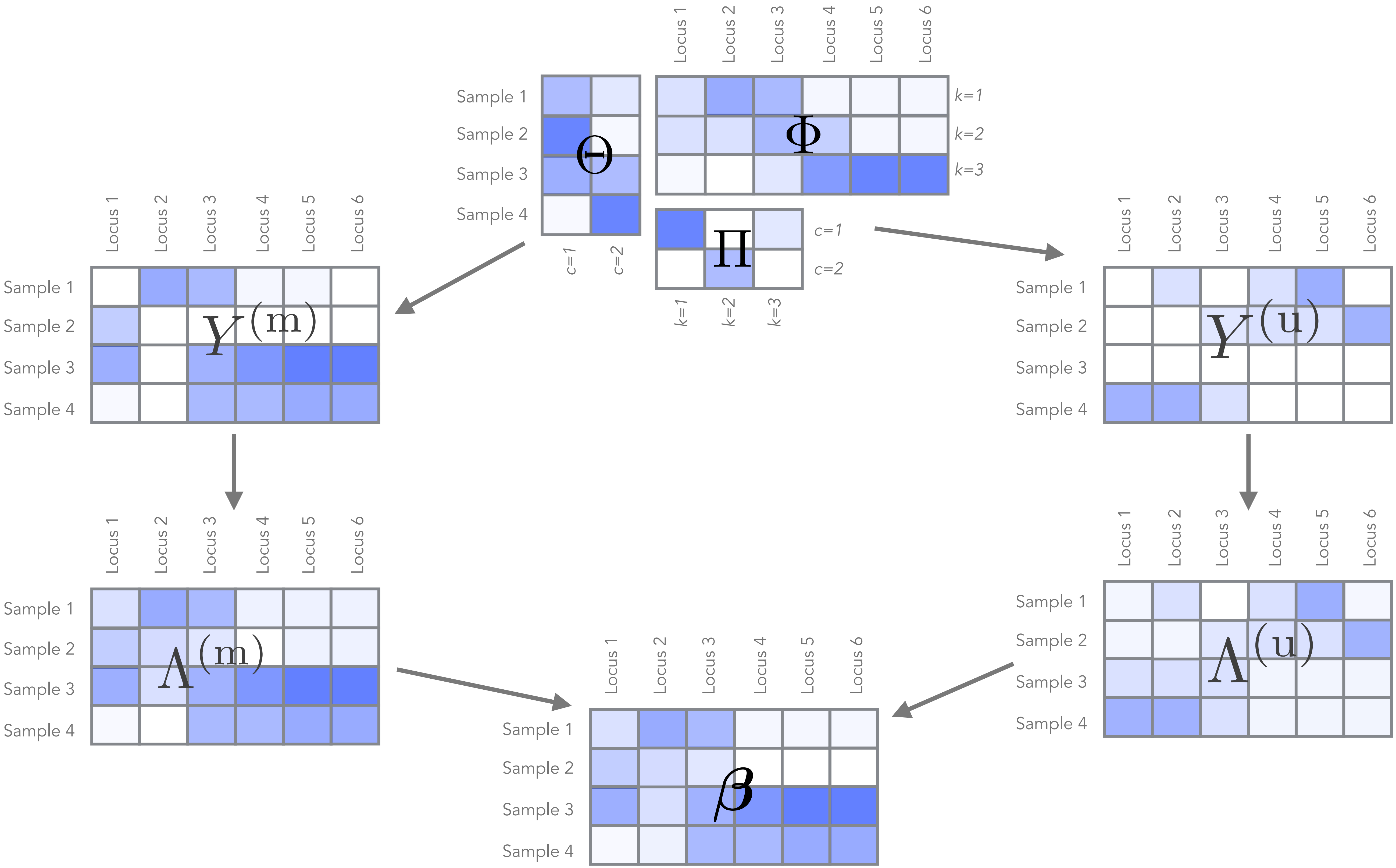
$$y_{ij}^{(\mathrm{m})} \sim \mathrm{Pois}\big(\gamma\, p_{ij}\big) \qquad y_{ij}^{(\mathrm{u})} \sim \mathrm{Pois}\big(\gamma\,(1 - p_{ij})\big)$$

$$\beta_{ij} \sim \mathrm{Beta}\left(b_0 + y_{ij}^{(\mathrm{m})},\, b_0 + y_{ij}^{(\mathrm{u})}\right)$$

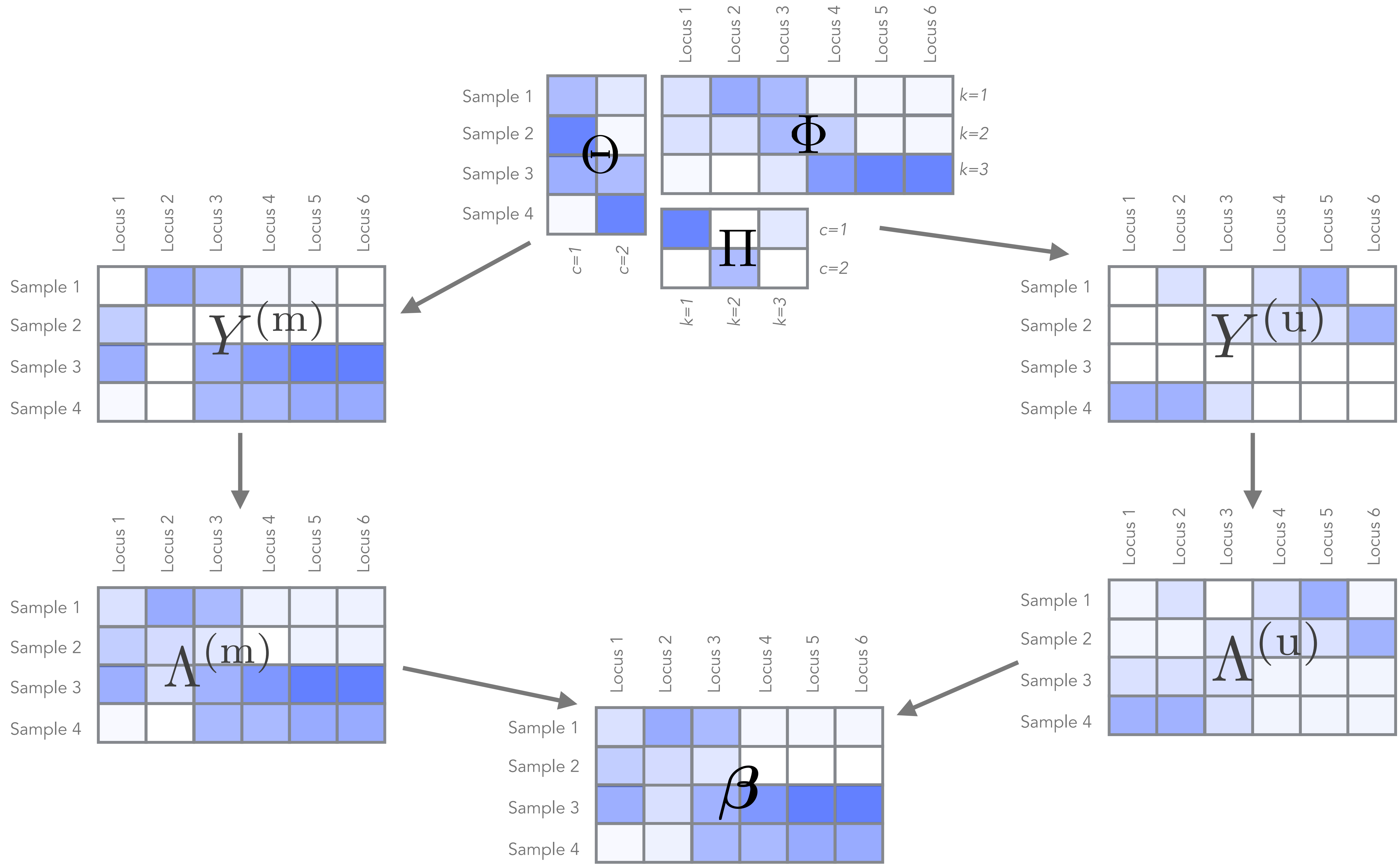
Beta Tucker decomposition



Beta Tucker decomposition

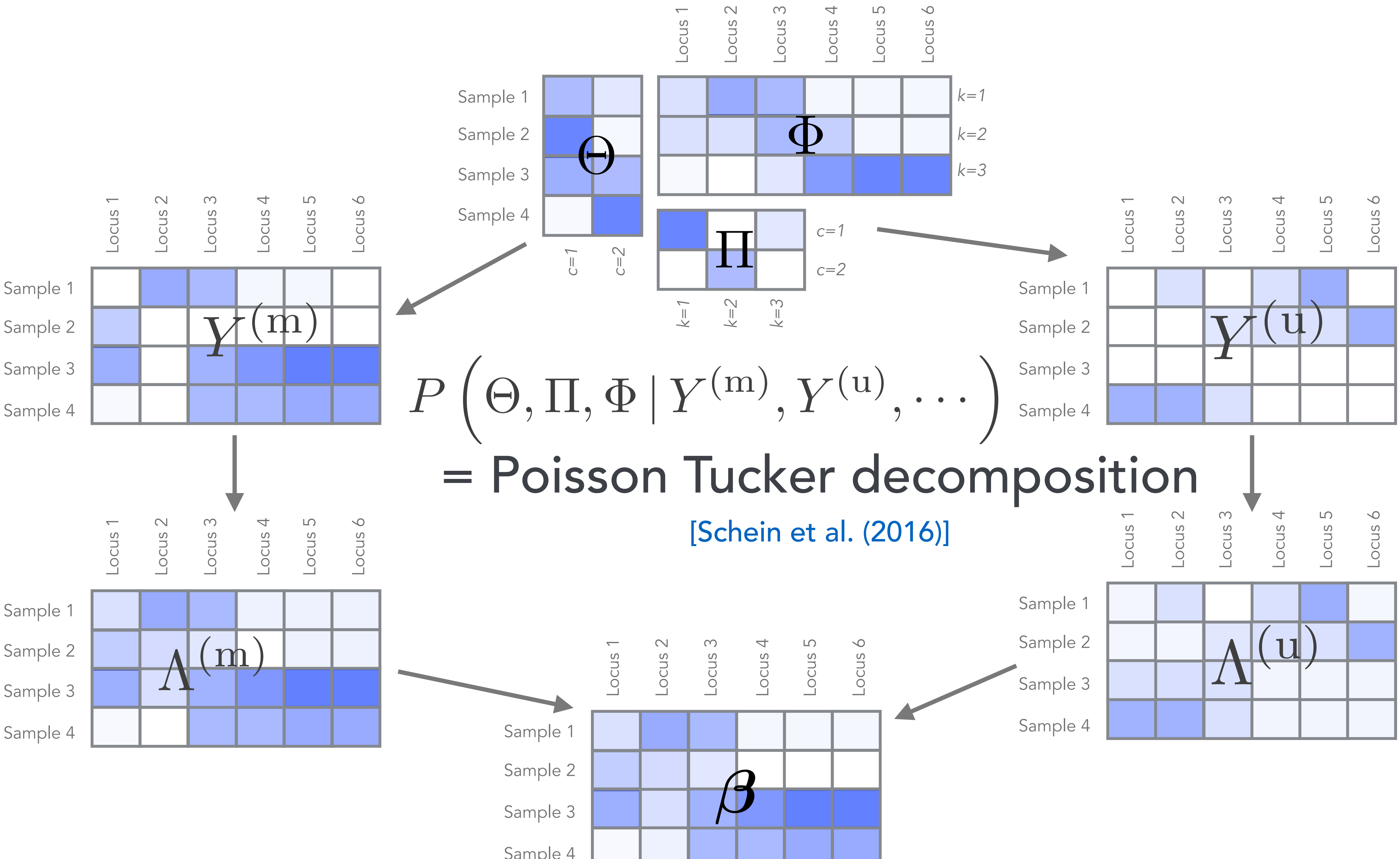


Inference



Locus 1
Locus 2
Locus 3
Locus 4
Locus 5
Locus 6
Sample 1
Sample 2
Sample 3
Sample 4
Θ
c=1
c=2
Φ
k=1
k=2
k=3
Π
$Y^{(m)}$
$Y^{(u)}$
$\Lambda^{(m)}$
$\Lambda^{(u)}$
$\boldsymbol{\beta}$

Inference

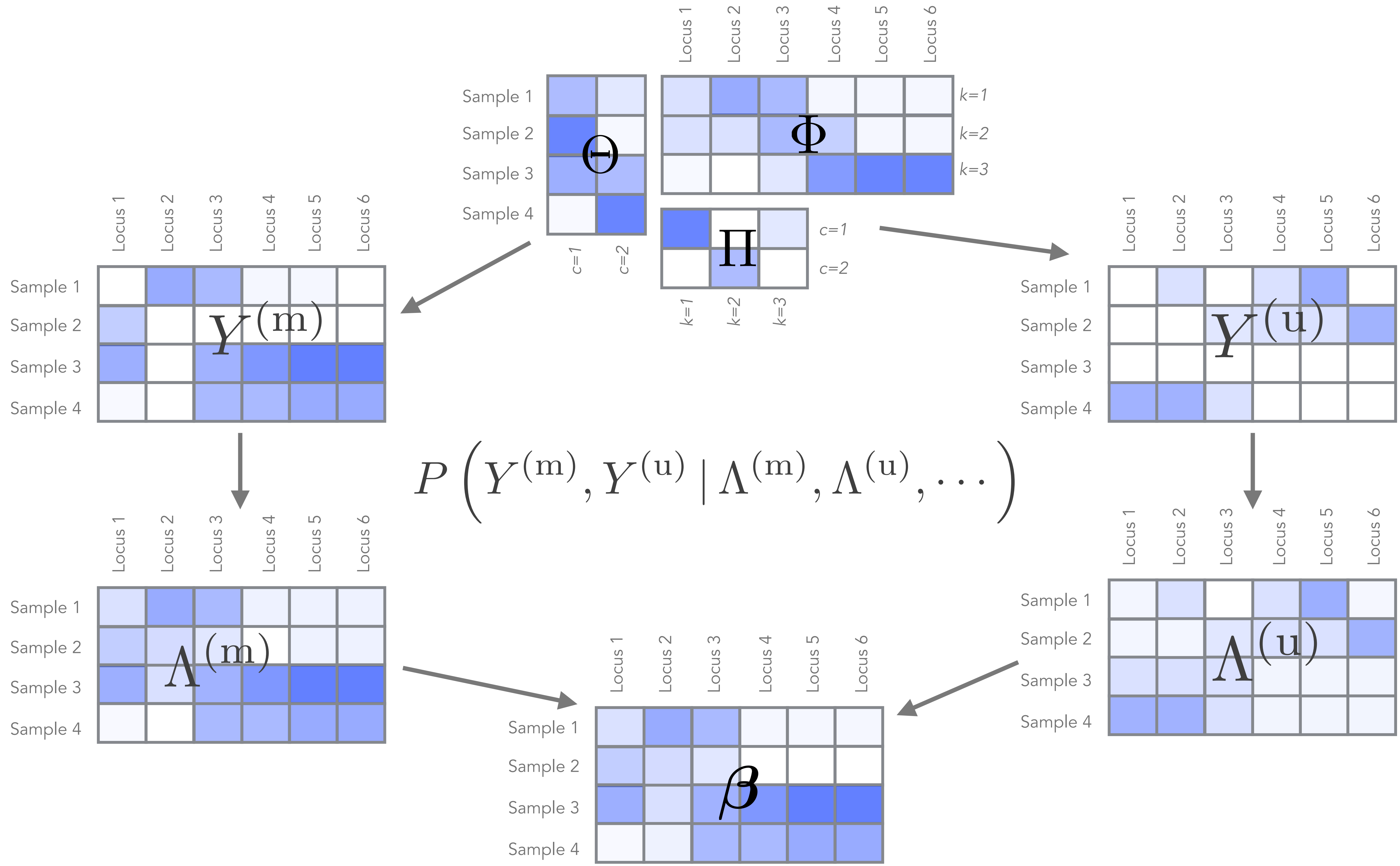


$$P\left(\Theta, \Pi, \Phi \,|\, Y^{(\mathrm{m})}, Y^{(\mathrm{u})}, \ldots\right)$$

= Poisson Tucker decomposition

[Schein et al. (2016)]

Inference


Locus 1
Locus 2
Locus 3
Locus 4
Locus 5
Locus 6
Sample 1
Sample 2
Sample 3
Sample 4
Θ
c=1
c=2
Φ
k=1
k=2
k=3
Π
$Y^{(\mathrm{m})}$
$Y^{(\mathrm{u})}$
$P\left(Y^{(\mathrm{m})}, Y^{(\mathrm{u})} \mid \Lambda^{(\mathrm{m})}, \Lambda^{(\mathrm{u})}, \ldots\right)$
$\Lambda^{(\mathrm{m})}$
$\Lambda^{(\mathrm{u})}$
$\boldsymbol{\beta}$

Inference

$$y_{ij}^{(\mathrm{m})} \sim \mathrm{Pois}\big(\gamma\, p_{ij}\big)$$

$$\lambda_{ij}^{(\mathrm{m})} \sim \mathrm{Gam}\left(b_0 + y_{ij}^{(\mathrm{m})},\ c_i\right)$$

Poisson is not conjugate to the gamma…

$$P(y_{ij}^{(\mathrm{m})} \mid \lambda_{ij}^{(\mathrm{m})}, \cdots) = ?$$

…but maybe the posterior
still has a closed form…

Inference

$$y_{ij}^{(\mathrm{m})} \sim \mathrm{Pois}\big(\gamma\, p_{ij}\big)$$

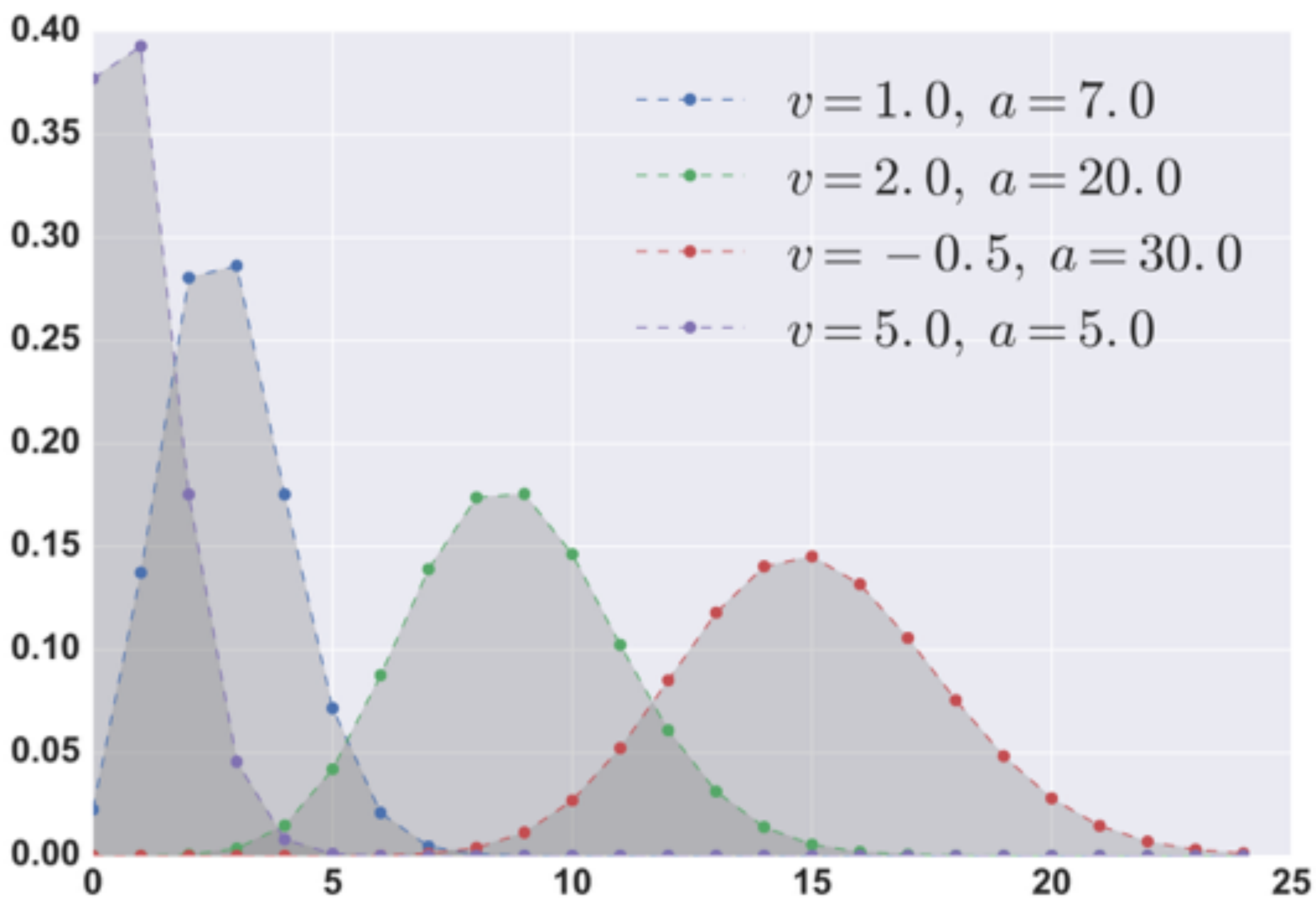
$$\lambda_{ij}^{(\mathrm{m})} \sim \mathrm{Gam}\left(b_0 + y_{ij}^{(\mathrm{m})},\, c_i\right)$$

The Bessel distribution! [Yuan & Kalbfleisch (2000)]

$$P(y_{ij}^{(\mathrm{m})} \mid \lambda_{ij}^{(\mathrm{m})}, \cdots) = \mathrm{Bes}\left(b_0 - 1,\, 2\sqrt{c_i \lambda_{ij}^{(\mathrm{m})}\, \gamma\, p_{ij}}\right)$$

The Bessel distribution

$$\mathrm{Bes}(y;\, v, a) \propto \frac{1}{y!\,\Gamma(y+v)} \left(\frac{a}{2}\right)^{2y+v}$$



Sampling the Bessel

It's easy and fast

[Amos (1974)]	Stable computation of Bessel functions
[Yuan & Kalbfleisch (2000)]	Basic properties of Bessel distribution
[Devroye (2002)]	Exact rejection sampling (four methods)
[Zhou (2015)]	Table sampling

https://github.com/aschein/fatwalrus

MCMC algorithm

$$P\left(Y^{(\mathrm{m})}, Y^{(\mathrm{u})} \mid \Lambda^{(\mathrm{m})}, \Lambda^{(\mathrm{u})}, \cdots\right) \qquad \mathcal{O}(2IJ)$$

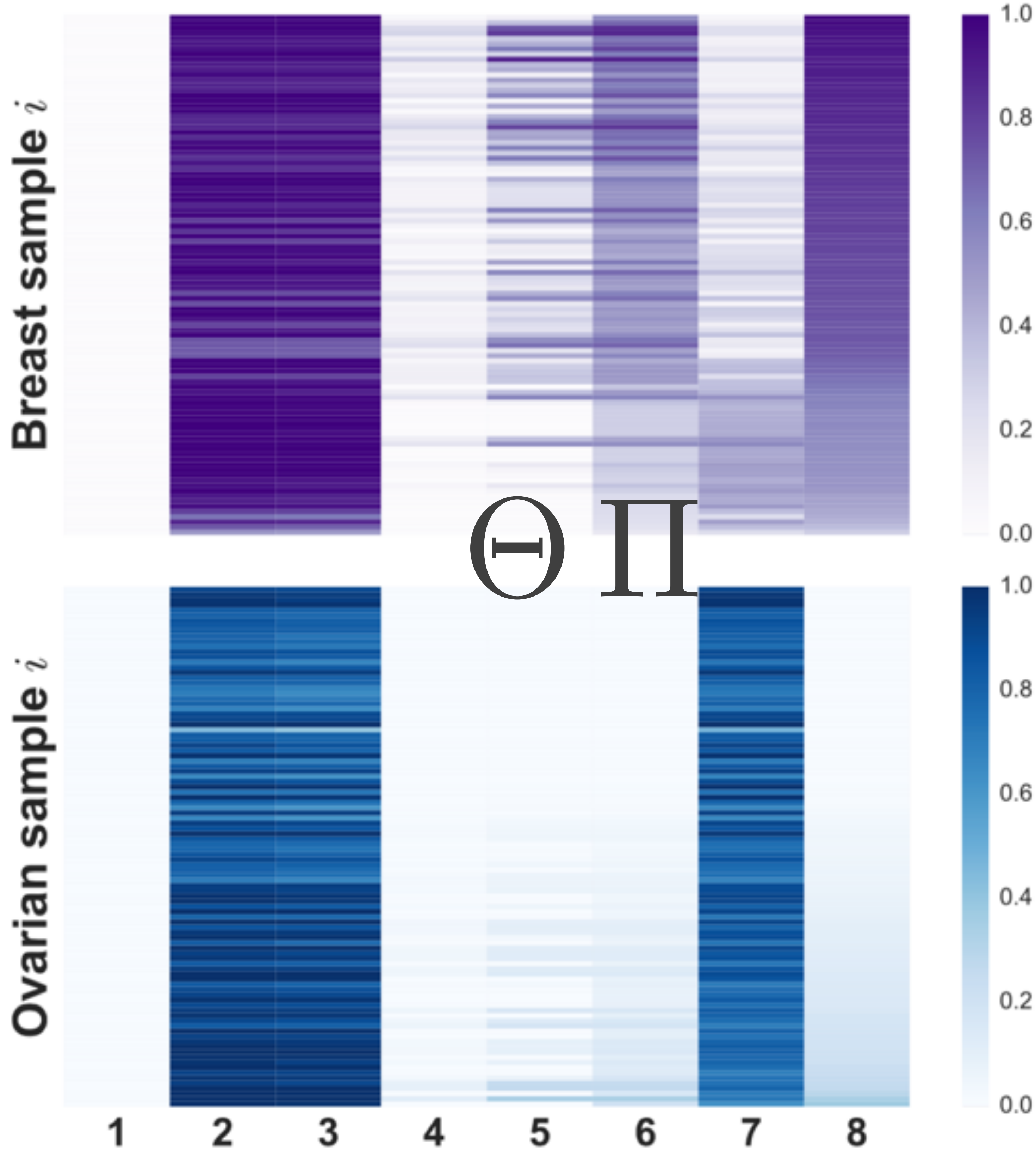
Sample Bessel counts

$$P\left(\Theta, \Pi, \Phi \mid Y^{(\mathrm{m})}, Y^{(\mathrm{u})}, \cdots\right) \qquad \mathcal{O}(CK|Y_{>0}|)$$

Poisson Tucker decomposition

γ controls sparsity!

Example results



Top locus in component 8 is in the promoter region of FLJ1030207

Hypomethylation of FLJ1030207 is a strong indicator of ovarian cancer [Model & Rujan (2009)]