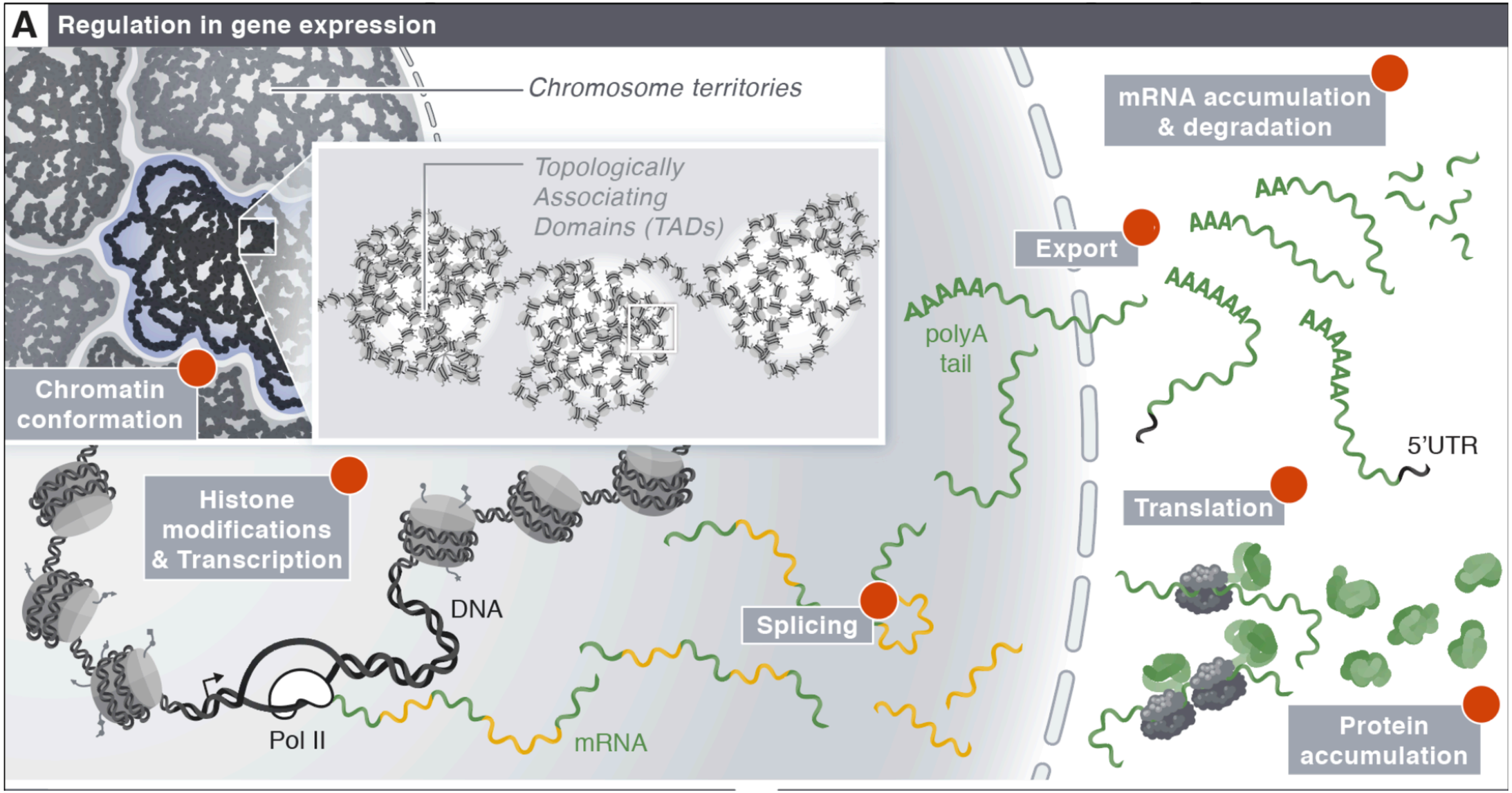
Greater than the sum of the parts: Learning relationships between histone modifications in single cells

Jake Yeung
Institute of Science and Technology Austria (ISTA)

PhD 2019 EPFL
HFSP Fellow 2019-2022 at Hubrecht, Sanger, ISTA
Incoming Group Leader (2024): Genentech

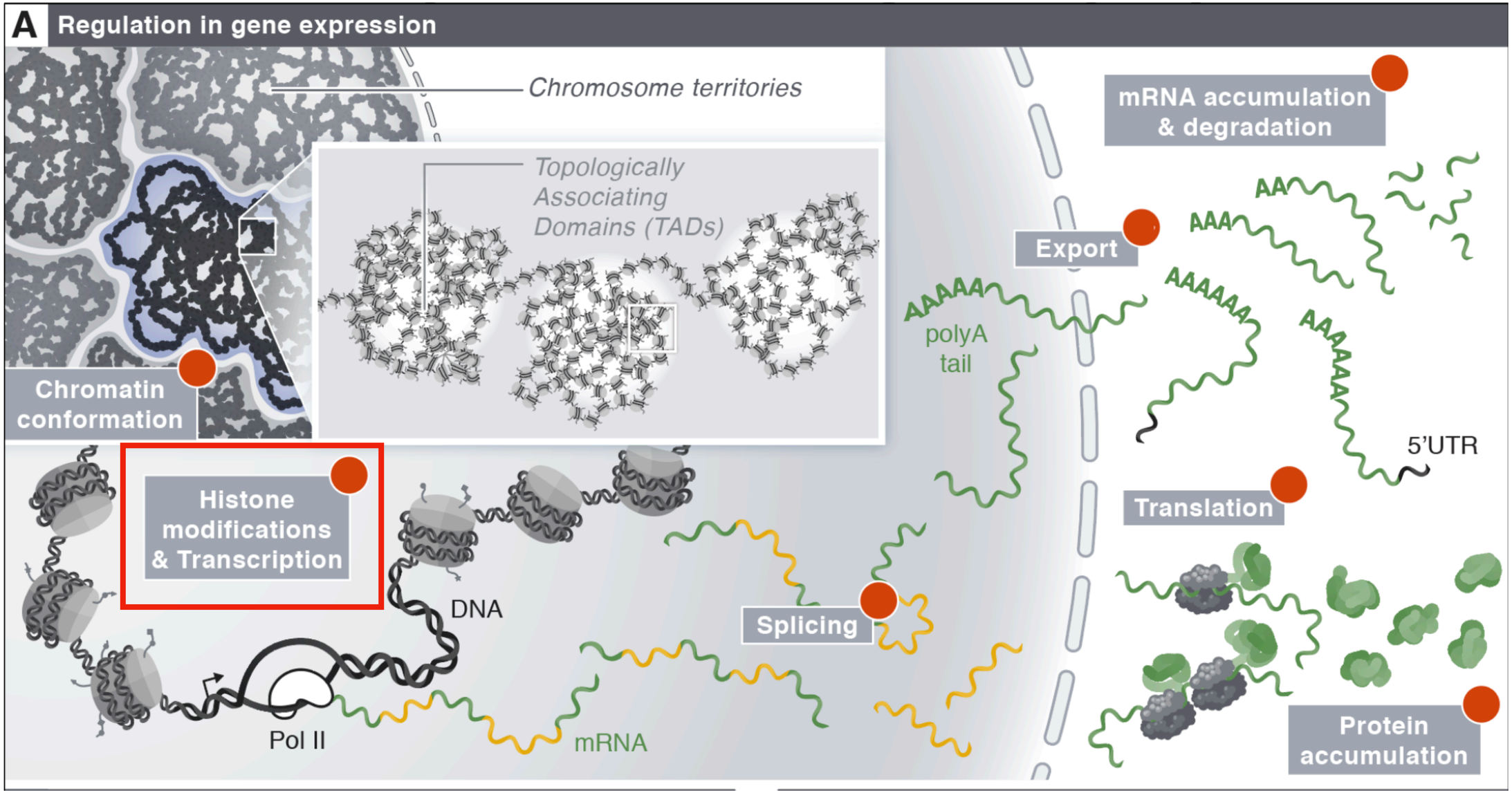
Monday July 3, 2023
BIRS : Data Science Challenges in Single-Cell Research

Single-cell epigenomics is increasingly multimodal, can we infer relationships between modalities?



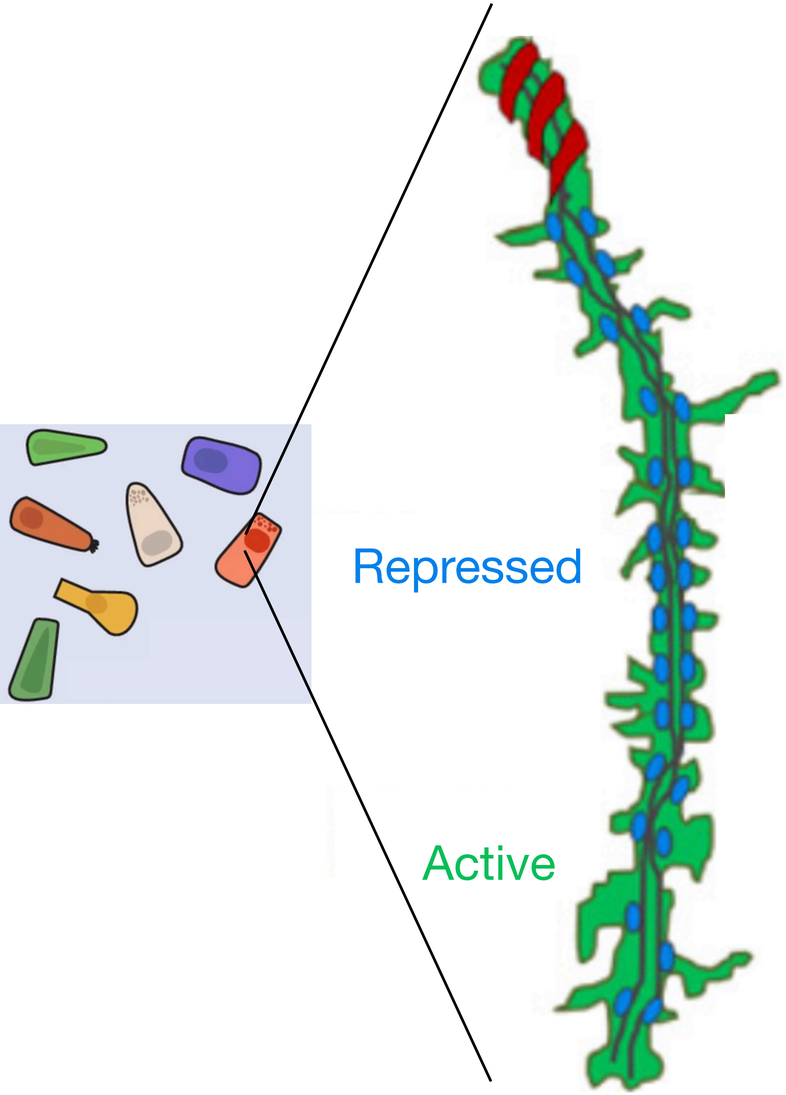
Adapted from Mermet, Yeung, Naef, 2017 *Cold Spring Harbor Perspectives*

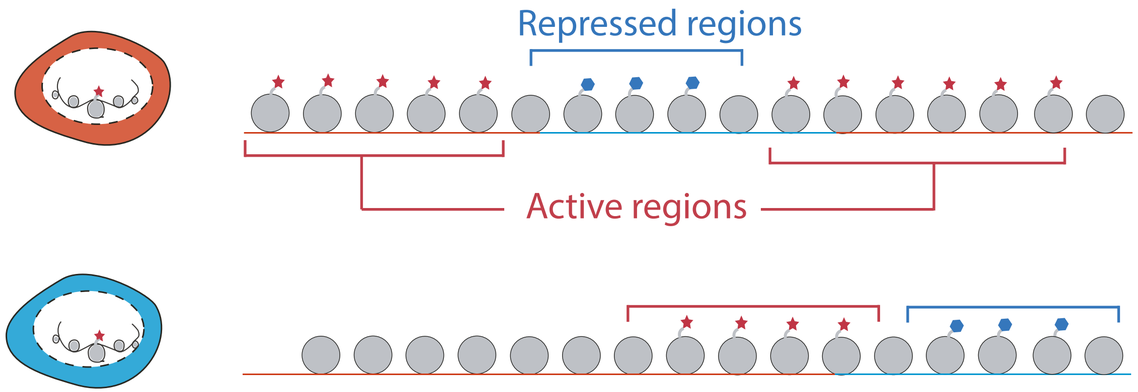
Single-cell epigenomics is increasingly multimodal, can we infer relationships between modalities?



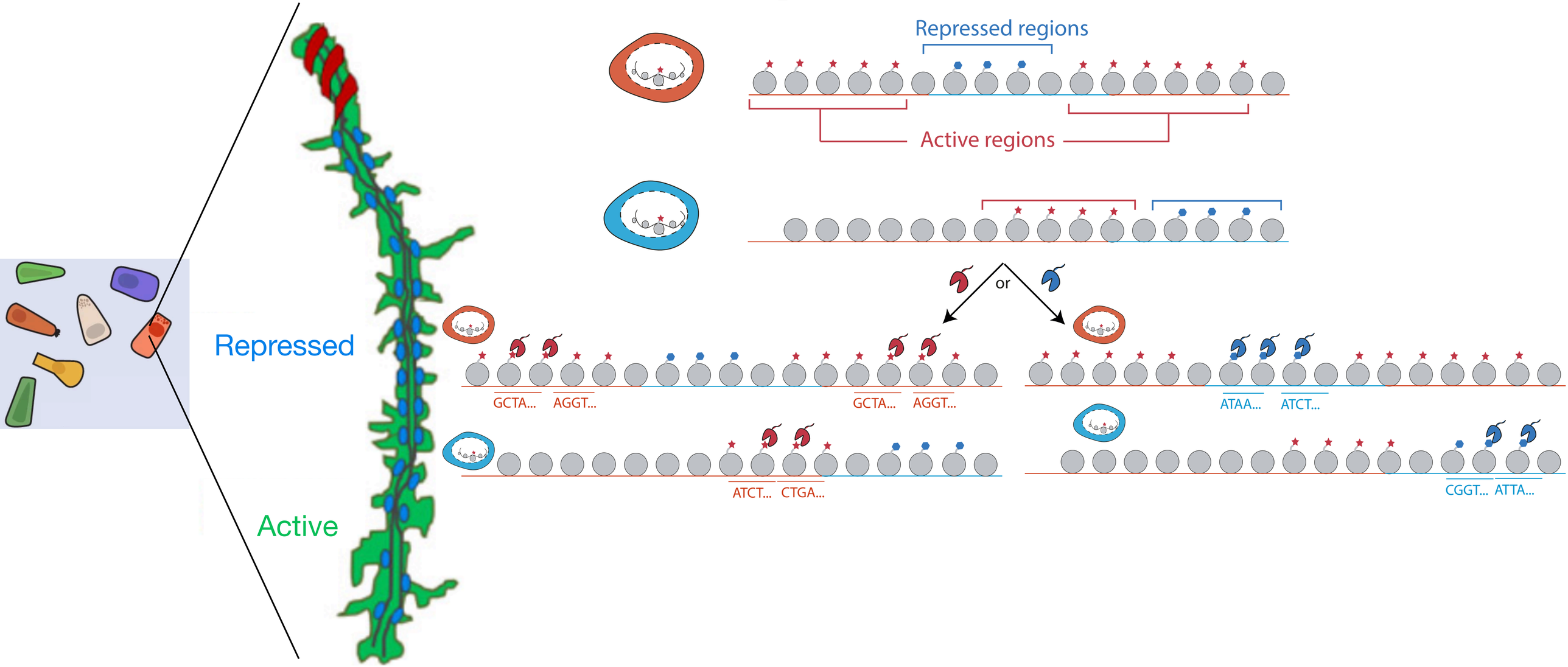
Adapted from Mermet, Yeung, Naef, 2017 *Cold Spring Harbor Perspectives*

Most chromatin profiling techniques measure only one histone modification per cell

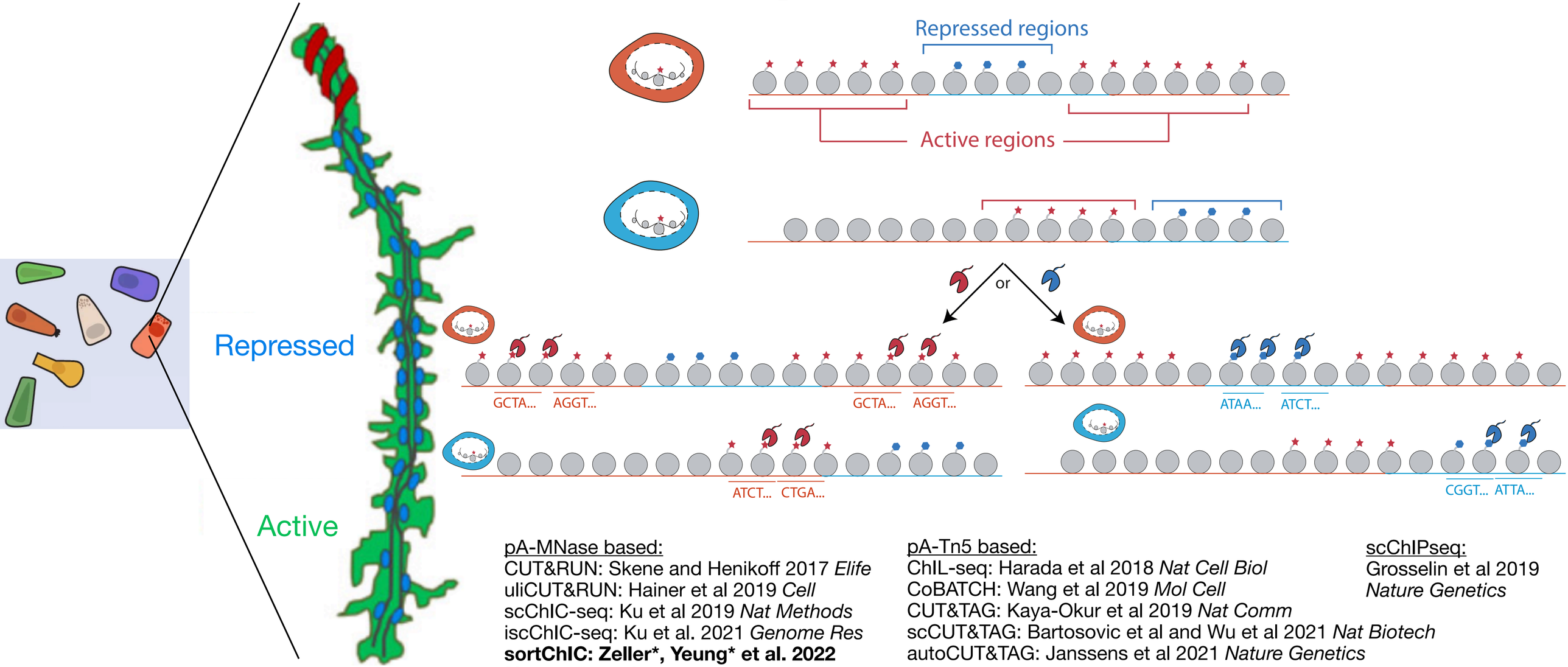




Most chromatin profiling techniques measure only one histone modification per cell



Most chromatin profiling techniques measure only one histone modification per cell



pA-MNase based:
CUT&RUN: Skene and Henikoff 2017 *Elife*
uliCUT&RUN: Hainer et al 2019 *Cell*
scChIC-seq: Ku et al 2019 *Nat Methods*
iscChIC-seq: Ku et al. 2021 *Genome Res*
sortChIC: Zeller*, Yeung* et al. 2022 *Nature Genetics*

pA-Tn5 based:
ChIL-seq: Harada et al 2018 *Nat Cell Biol*
CoBATCH: Wang et al 2019 *Mol Cell*
CUT&TAG: Kaya-Okur et al 2019 *Nat Comm*
scCUT&TAG: Bartosovic et al and Wu et al 2021 *Nat Biotech*
autoCUT&TAG: Janssens et al 2021 *Nature Genetics*

scChIPseq:
Grosselin et al 2019 *Nature Genetics*

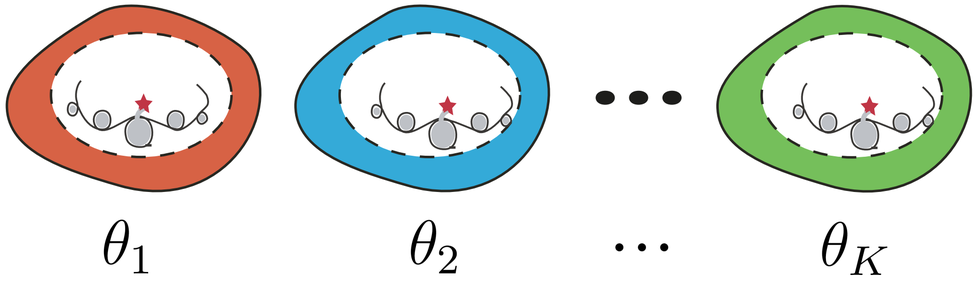
For single-modality analysis, latent Dirichlet allocation (LDA) is a natural way to model sparse counts

To generate a cut location $w_{d,n}$ in cell d for the nth read:

1) Choose a latent variable (topic)

$$z_{d,n} \sim \text{Multinomial}\left(1, \vec{\theta}_d\right)$$

Latent factors



See also:
Structure from Pritchard, Stephens, Donnelly 2000
LDA from Blei, Ng, Jordan 2003

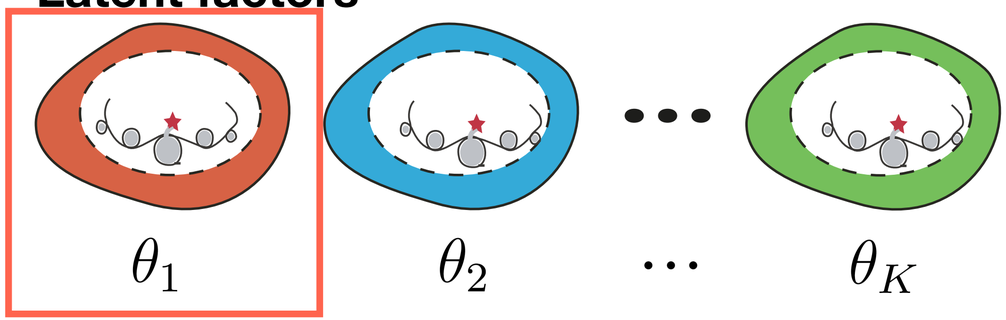
For single-modality analysis, latent Dirichlet allocation (LDA) is a natural way to model sparse counts

To generate a cut location $w_{d,n}$ in cell d for the nth read:

1) Choose a latent variable (topic)

$$z_{d,n} \sim \text{Multinomial}\left(1, \vec{\theta}_d\right)$$

Latent factors



See also:
Structure from Pritchard, Stephens, Donnelly 2000
LDA from Blei, Ng, Jordan 2003

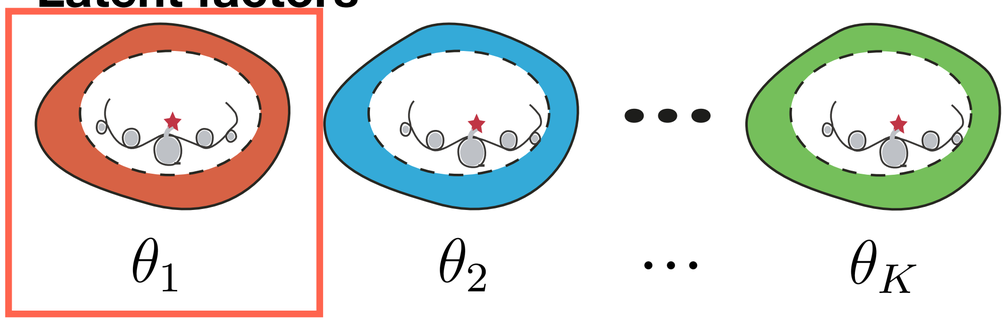
For single-modality analysis, latent Dirichlet allocation (LDA) is a natural way to model sparse counts

To generate a cut location $w_{d,n}$ in cell d for the nth read:

1) Choose a latent variable (topic)

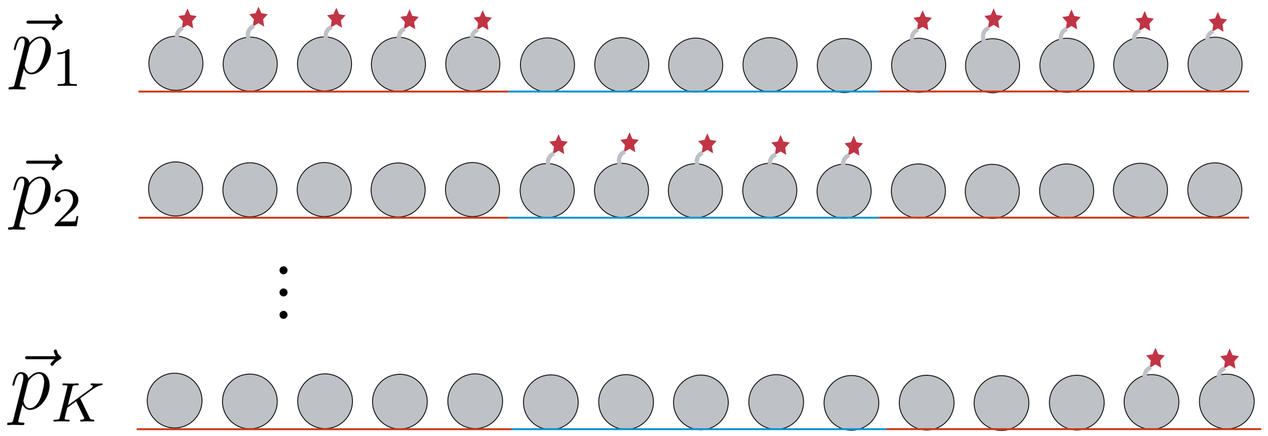
$$z_{d,n} \sim \text{Multinomial}\left(1, \vec{\theta}_d\right)$$

Latent factors



2) Choose a genomic region, given the latent variable

$$w_{d,n} \sim \text{Multinomial}\left(1, \vec{p}_{z_{d,n}}\right)$$



See also:
Structure from Pritchard, Stephens, Donnelly 2000
LDA from Blei, Ng, Jordan 2003

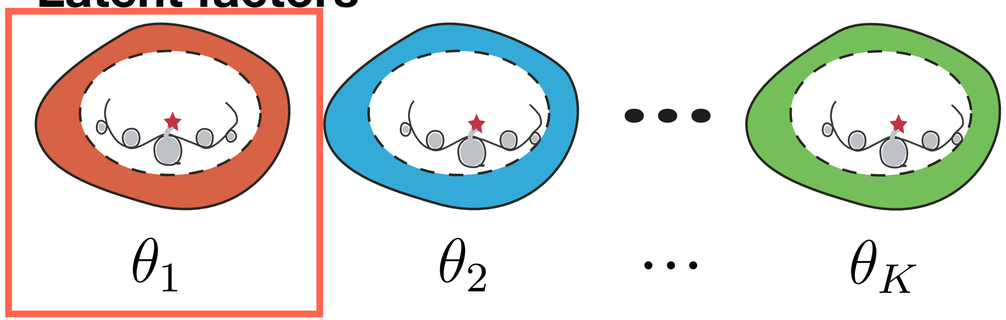
For single-modality analysis, latent Dirichlet allocation (LDA) is a natural way to model sparse counts

To generate a cut location $w_{d,n}$ in cell d for the nth read:

1) Choose a latent variable (topic)

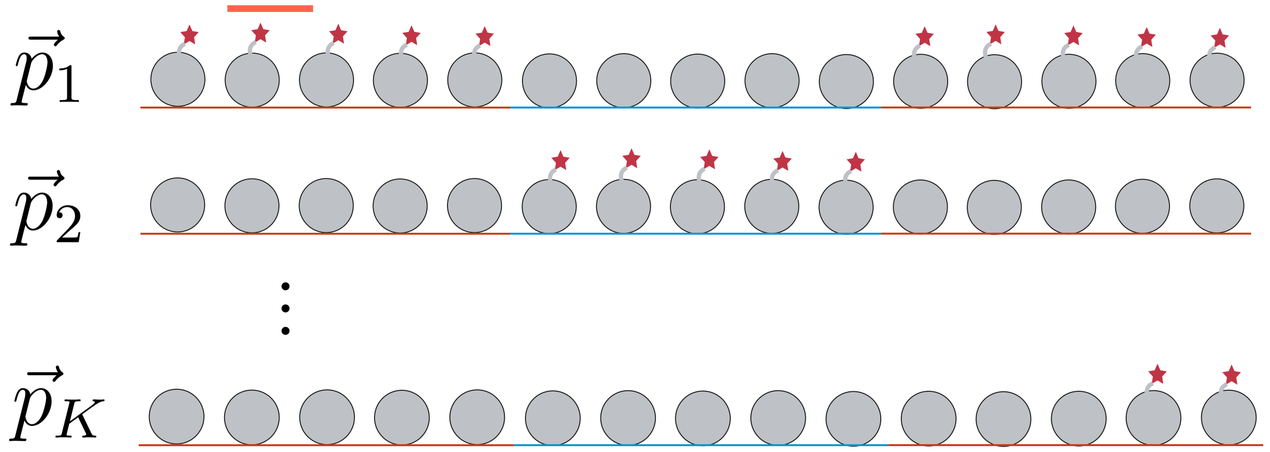
$$z_{d,n} \sim \text{Multinomial}\left(1, \vec{\theta}_d\right)$$

Latent factors



2) Choose a genomic region, given the latent variable

$$w_{d,n} \sim \text{Multinomial}\left(1, \vec{p}_{z_{d,n}}\right)$$



See also:
Structure from Pritchard, Stephens, Donnelly 2000
LDA from Blei, Ng, Jordan 2003

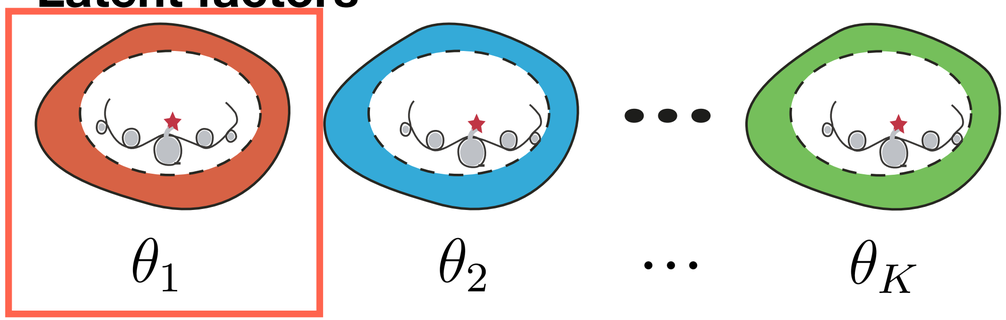
For single-modality analysis, latent Dirichlet allocation (LDA) is a natural way to model sparse counts

To generate a cut location $w_{d,n}$ in cell d for the nth read:

1) Choose a latent variable (topic)

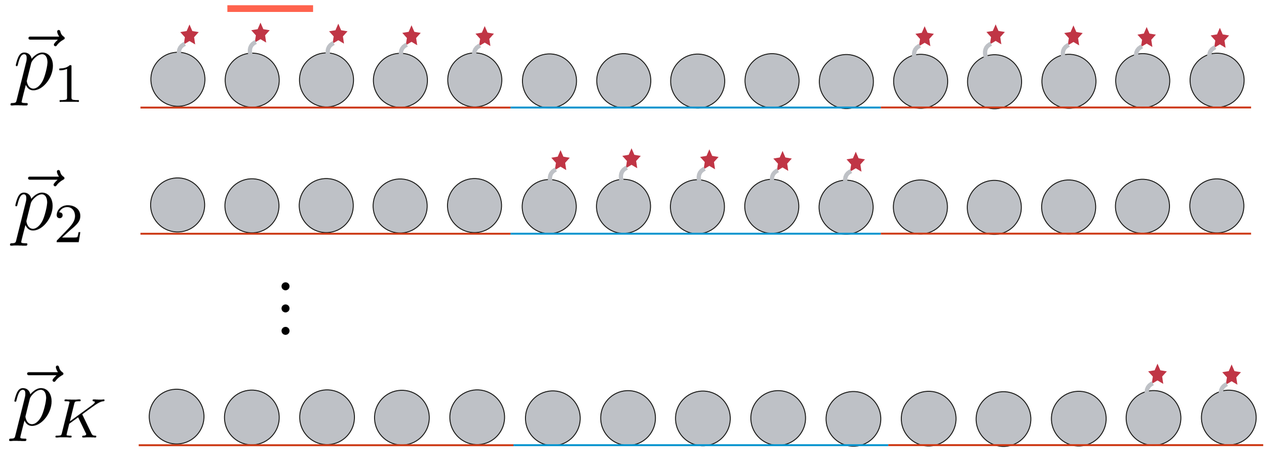
$$\vec{\theta}_d \sim \text{Dirichlet}\,(\alpha)$$
$$z_{d,n} \sim \text{Multinomial}\left(1, \vec{\theta}_d\right)$$

Latent factors



2) Choose a genomic region, given the latent variable

$$\vec{p}_k \sim \text{Dirichlet}(\lambda)$$
$$w_{d,n} \sim \text{Multinomial}\left(1, \vec{p}_{z_{d,n}}\right)$$



See also:
Structure from Pritchard, Stephens, Donnelly 2000
LDA from Blei, Ng, Jordan 2003

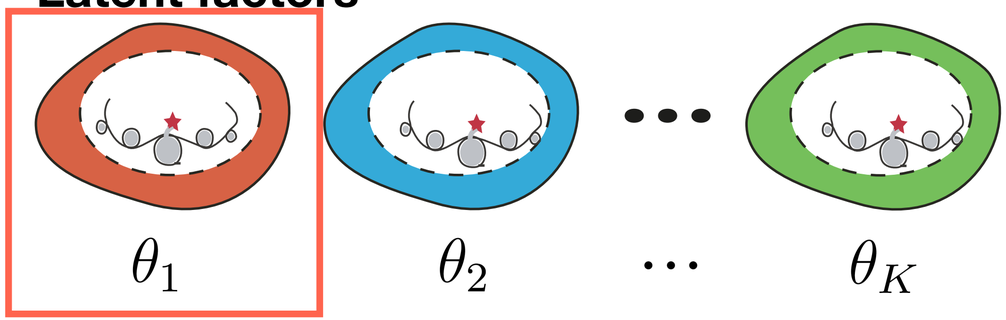
For single-modality analysis, latent Dirichlet allocation (LDA) is a natural way to model sparse counts

To generate a cut location $w_{d,n}$ in cell d for the nth read:

1) Choose a latent variable (topic)

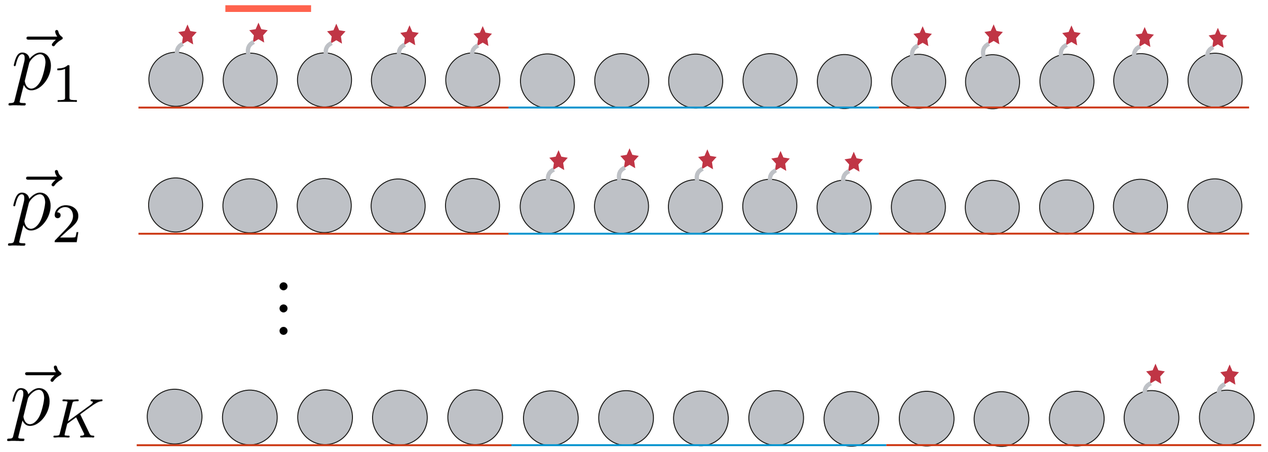
$$\vec{\theta}_d \sim \text{Dirichlet}\,(\alpha)$$
$$z_{d,n} \sim \text{Multinomial}\left(1, \vec{\theta}_d\right)$$

Latent factors



2) Choose a genomic region, given the latent variable

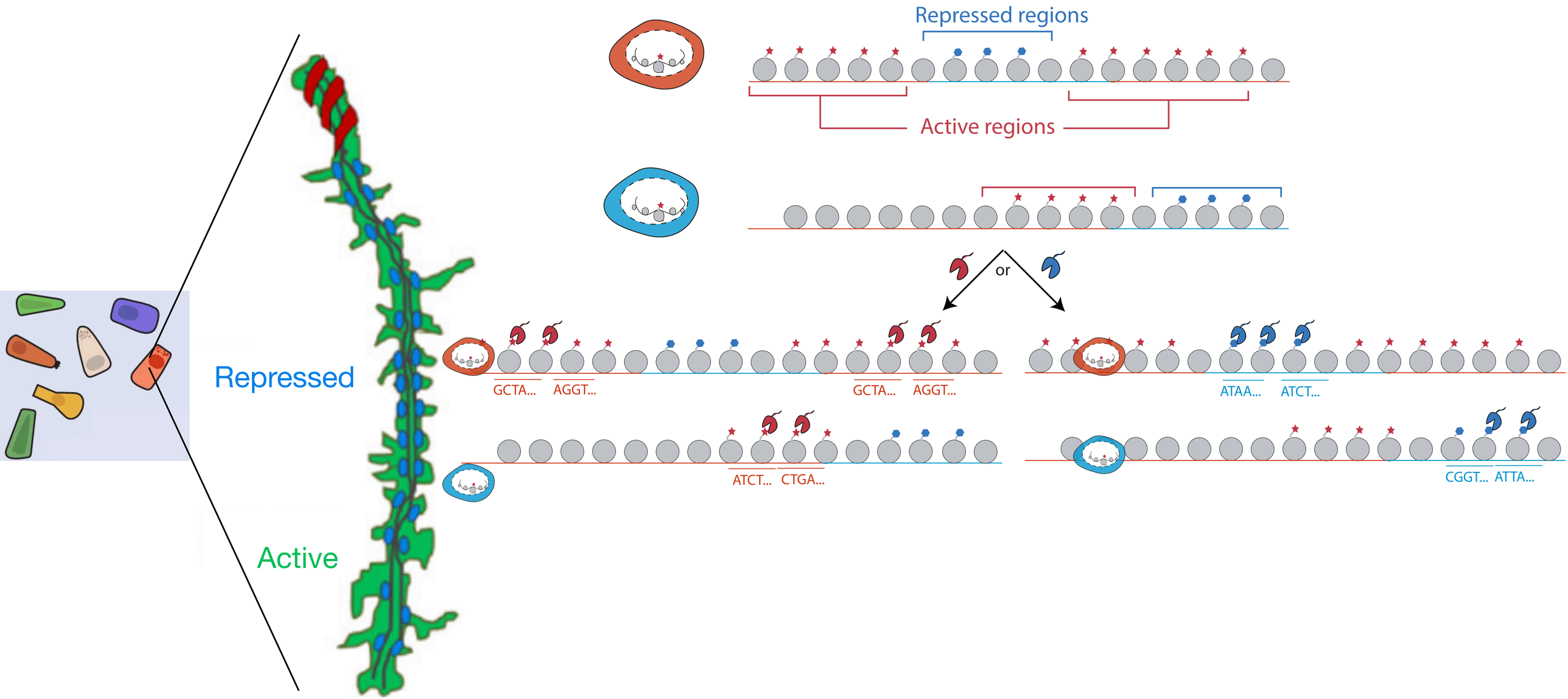
$$\vec{p}_k \sim \text{Dirichlet}(\lambda)$$
$$w_{d,n} \sim \text{Multinomial}\left(1, \vec{p}_{z_{d,n}}\right)$$



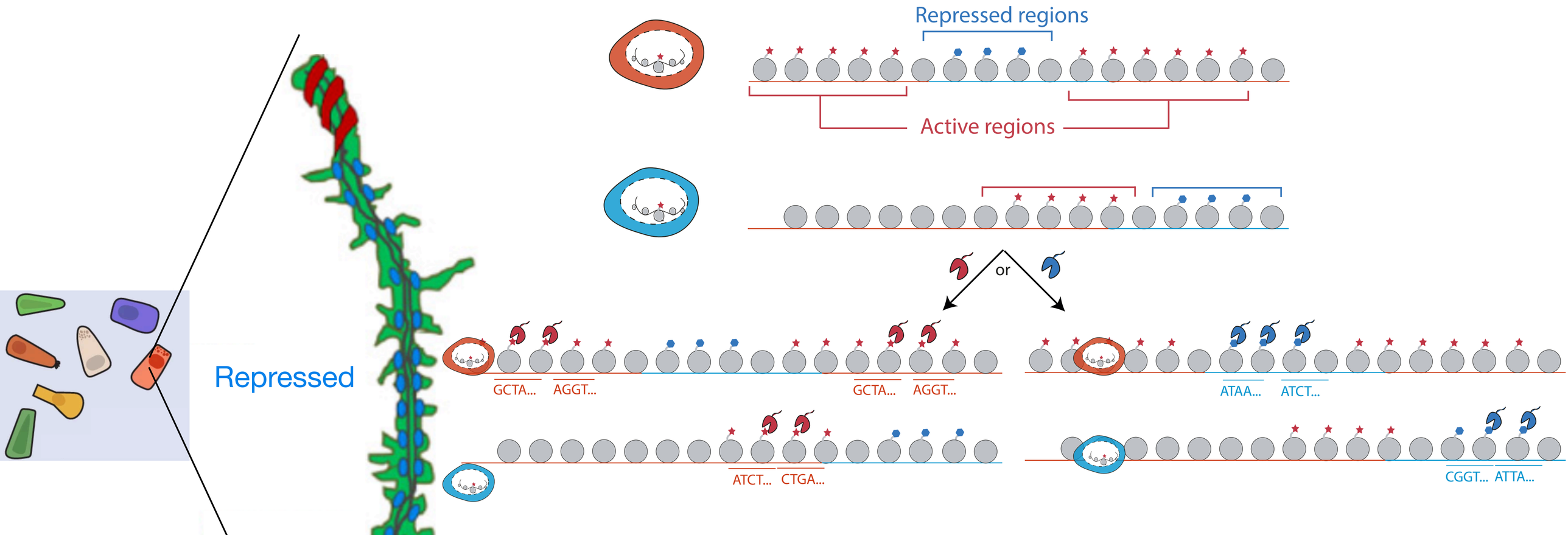
Parameters **θ** and **P** are inferred by collapsed Gibbs sampling

See also:
Structure from Pritchard, Stephens, Donnelly 2000
LDA from Blei, Ng, Jordan 2003

Can we go beyond one histone mark per cell?

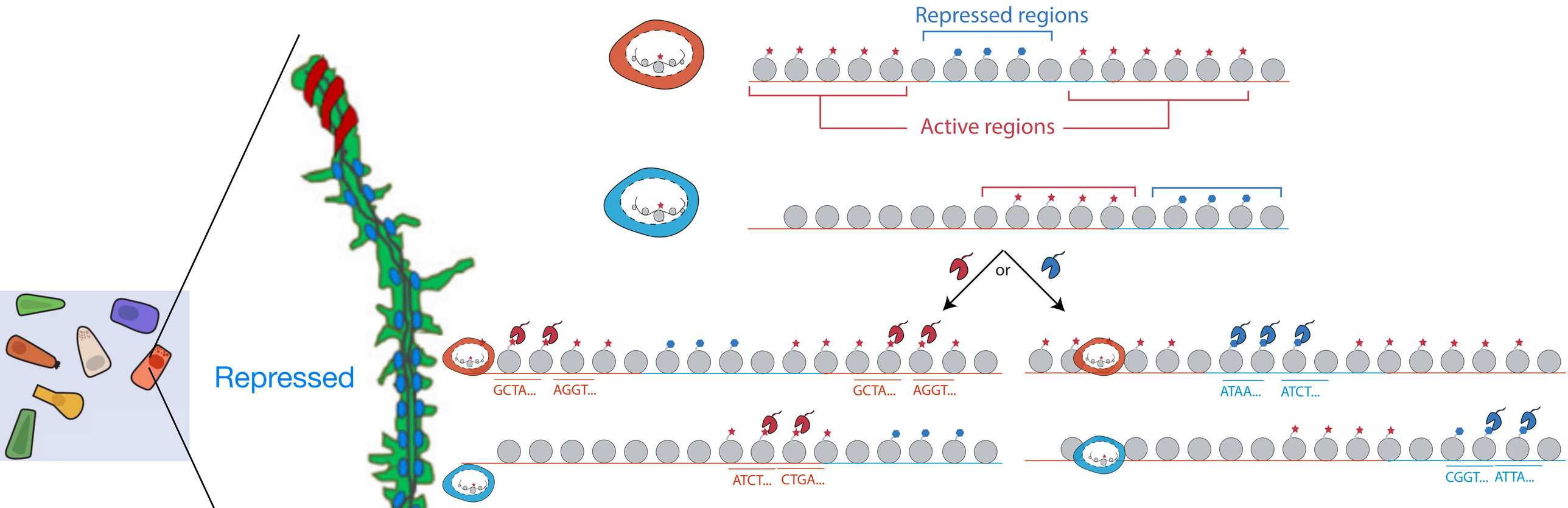


Can we go beyond one histone mark per cell?



Experimental solutions (barcoding):
Gopalan et al 2021 *Mol Cell*
Meers et al 2022 *Nat Biotech*
Bartosovic et al 2022 *Nat Biotech*
Stuart et al 2022 *Nat Biotech*
Lochs et al 2023 bioRxiv

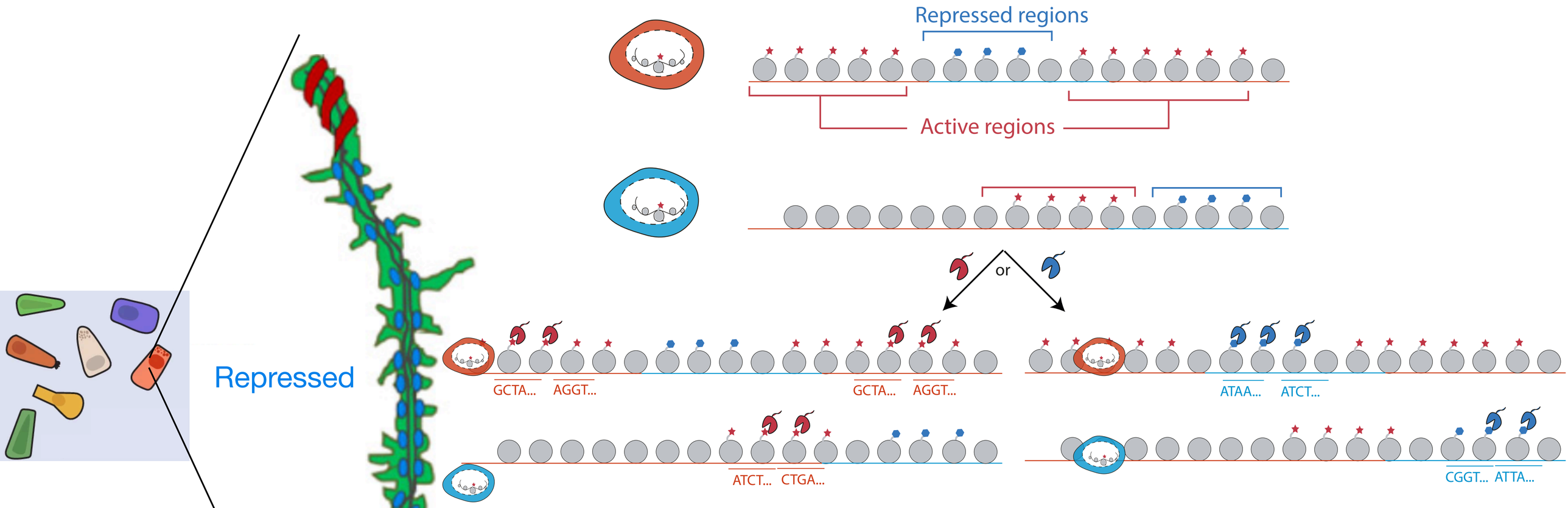
Can we go beyond one histone mark per cell?



Experimental solutions (barcoding):
Gopalan et al 2021 *Mol Cell*
Meers et al 2022 *Nat Biotech*
Bartosovic et al 2022 *Nat Biotech*
Stuart et al 2022 *Nat Biotech*
Lochs et al 2023 bioRxiv

Data science solution (mix and deconvolve):
Yeung*, Florescu* et al *Nat Biotech* 2023

Can we go beyond one histone mark per cell?



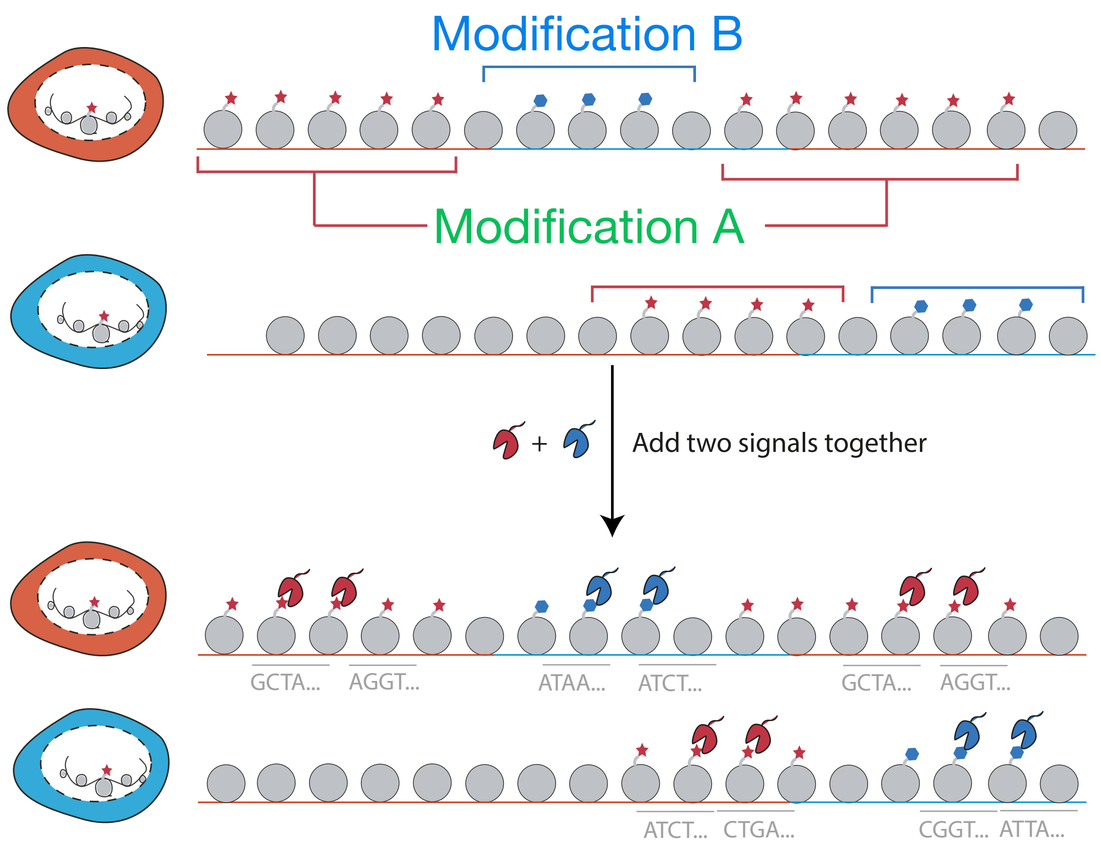
Experimental solutions (barcoding):
Gopalan et al 2021 *Mol Cell*
Meers et al 2022 *Nat Biotech*
Bartosovic et al 2022 *Nat Biotech*
Stuart et al 2022 *Nat Biotech*
Lochs et al 2023 bioRxiv

Data science solution (mix and deconvolve):

Yeung*, Florescu* et al *Nat Biotech* 2023

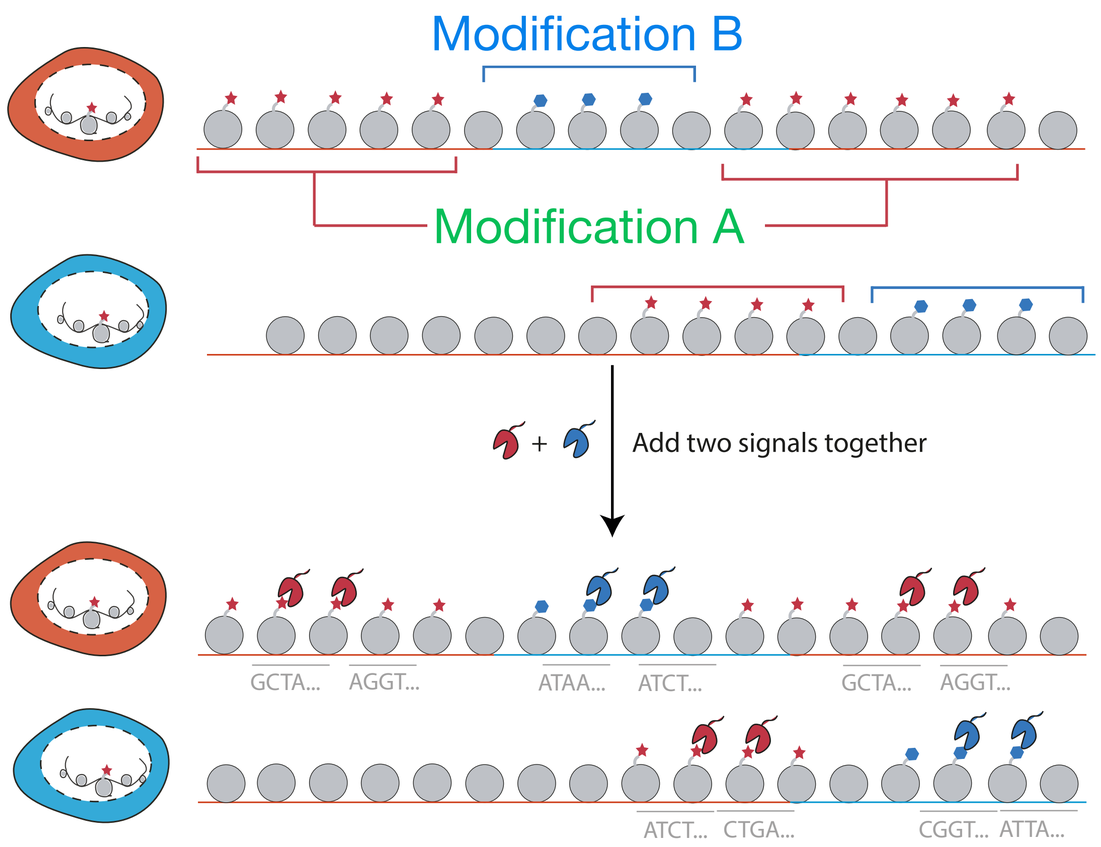
How can we extract unique insights from this multimodal data?

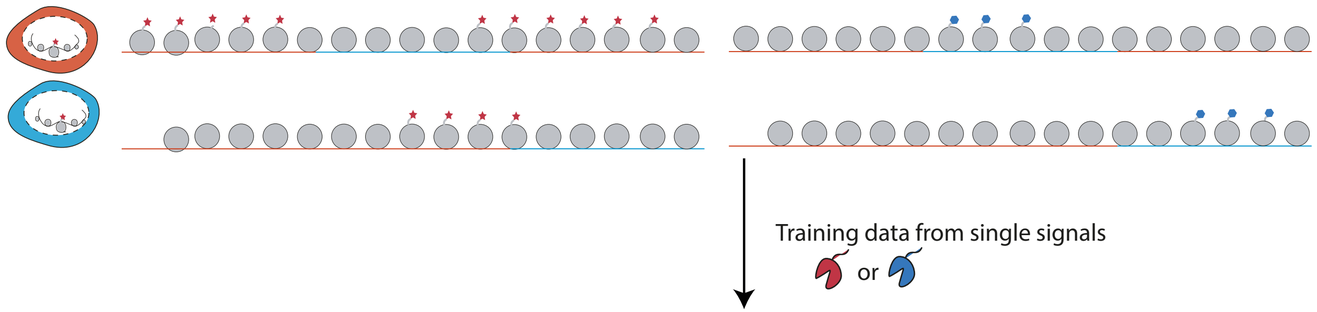
scChIX-seq multiplexes two histone modifications together, then deconvolves the mixed signal



scChIX-seq: chromatin immunocleavage and unmixing
Yeung*, Florescu* et al *Nature Biotech* 2023

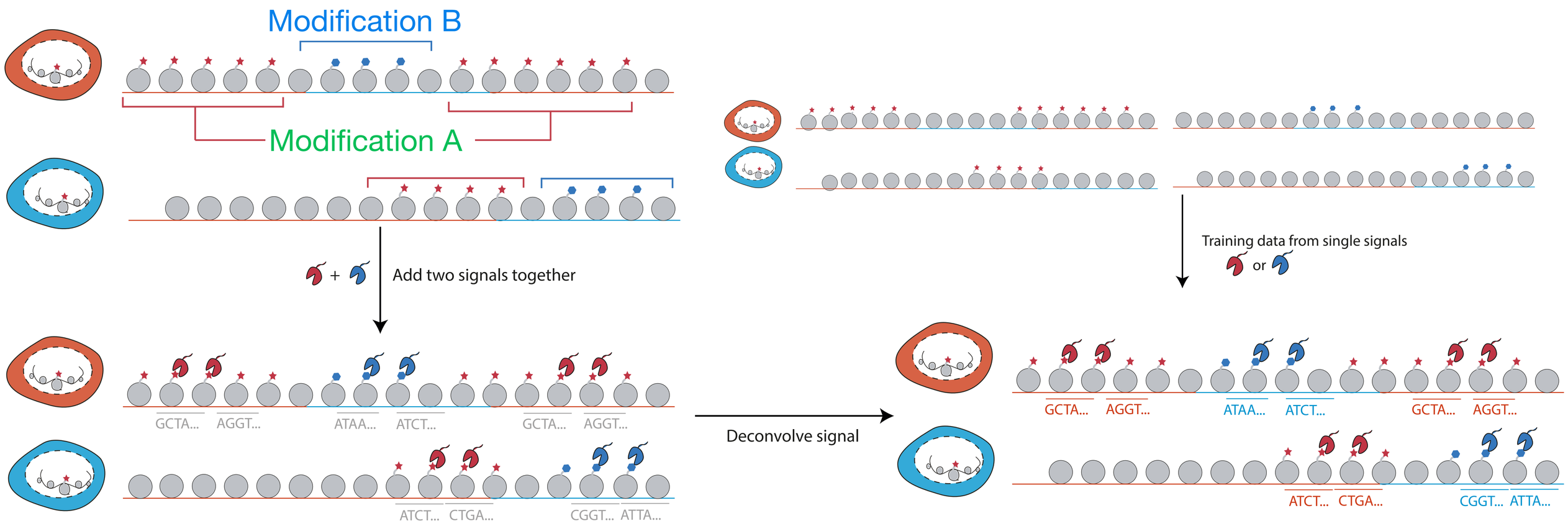
scChIX-seq multiplexes two histone modifications together, then deconvolves the mixed signal





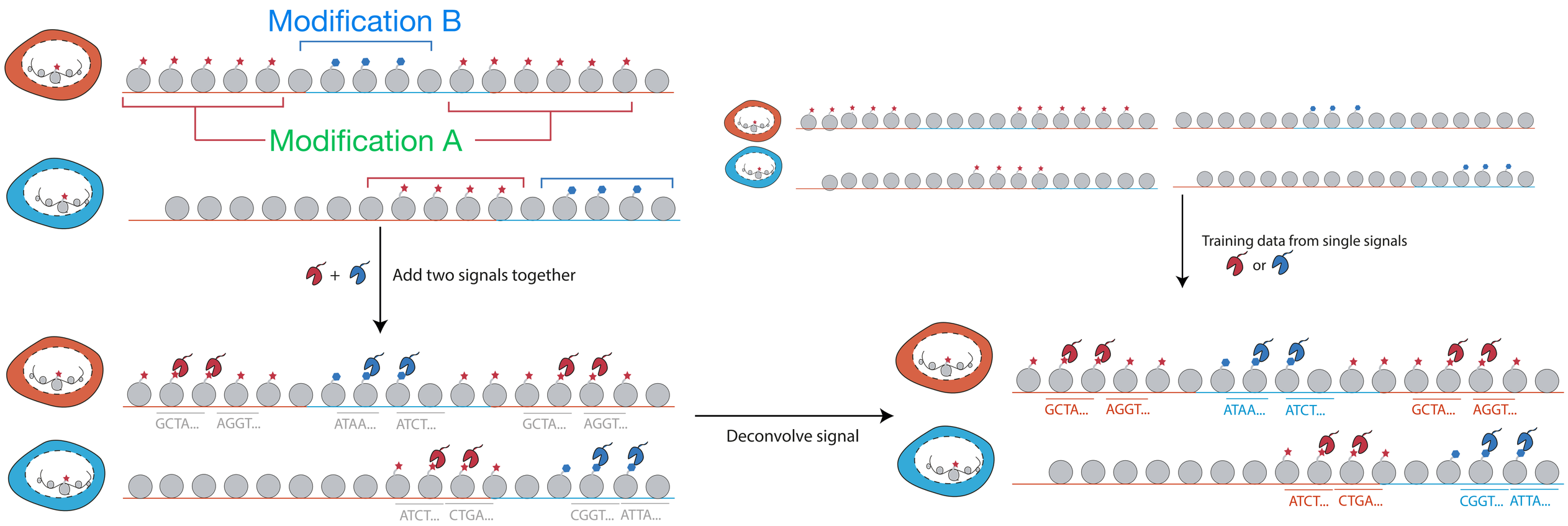
scChIX-seq: chromatin immunocleavage and unmixing
Yeung*, Florescu* et al *Nature Biotech* 2023

scChIX-seq multiplexes two histone modifications together, then deconvolves the mixed signal



scChIX-seq: chromatin immunocleavage and unmixing
Yeung*, Florescu* et al *Nature Biotech* 2023

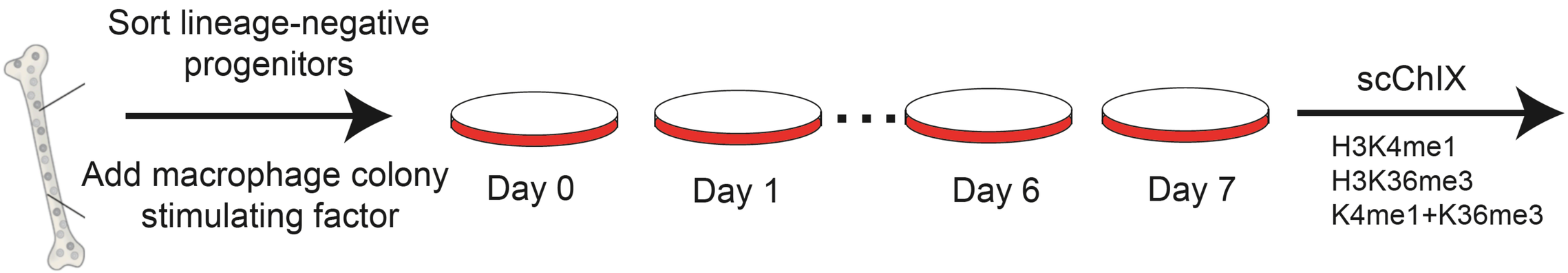
scChIX-seq multiplexes two histone modifications together, then deconvolves the mixed signal



scChIX-seq: chromatin immunocleavage and unmixing
Yeung*, Florescu* et al *Nature Biotech* 2023

Extending multinomial models to allow linear combinations of profiles

Apply scChIX-seq to uncover dynamic relationships between two active histone marks



Experimentalists:

Maria Florescu

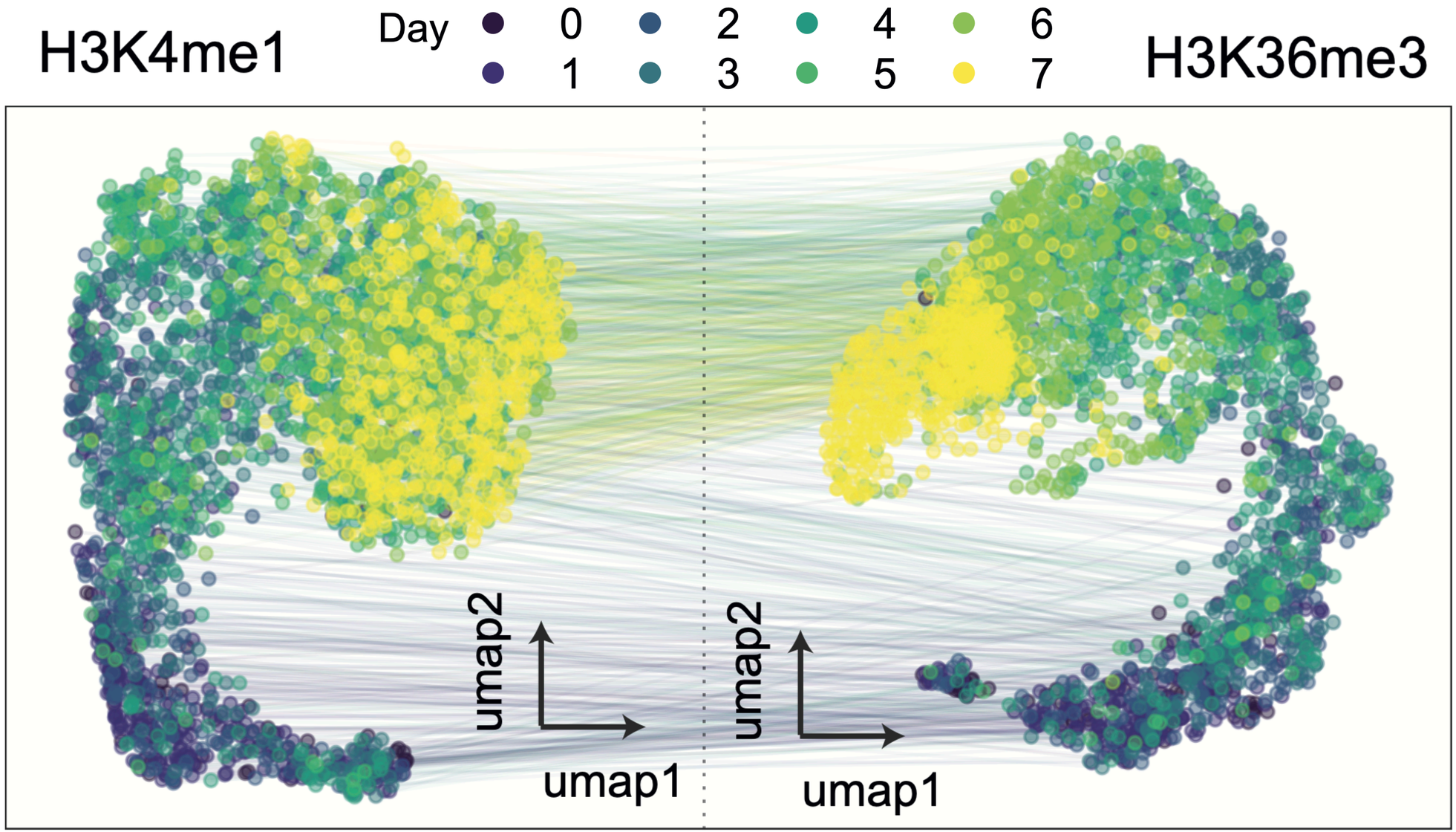
Max Wellenstein

Peter Zeller

Alexander van Oudenaarden group

H3K4me1: active and primed regions
H3K36me3: transcription

scChIX-seq connects H3K4me1 and H3K36me3 dynamics in single cells



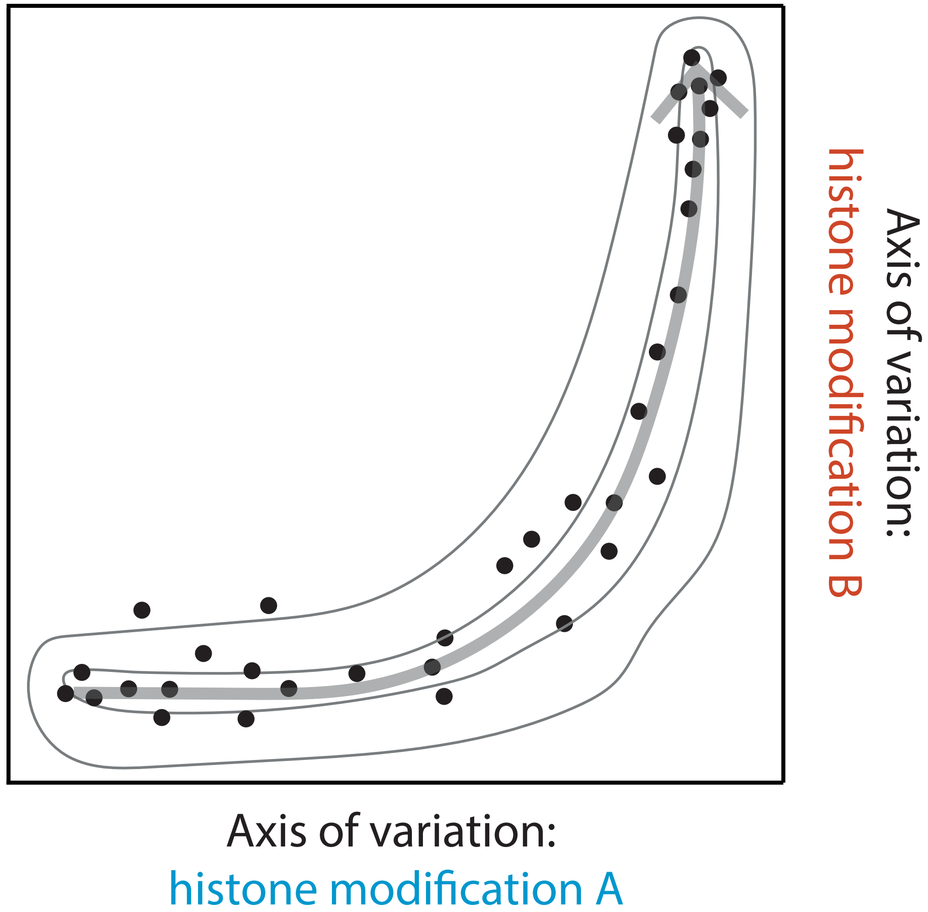
UMAP of cell-cell relationship matrix

scChIX-seq can uncover relationships *between* histone modifications

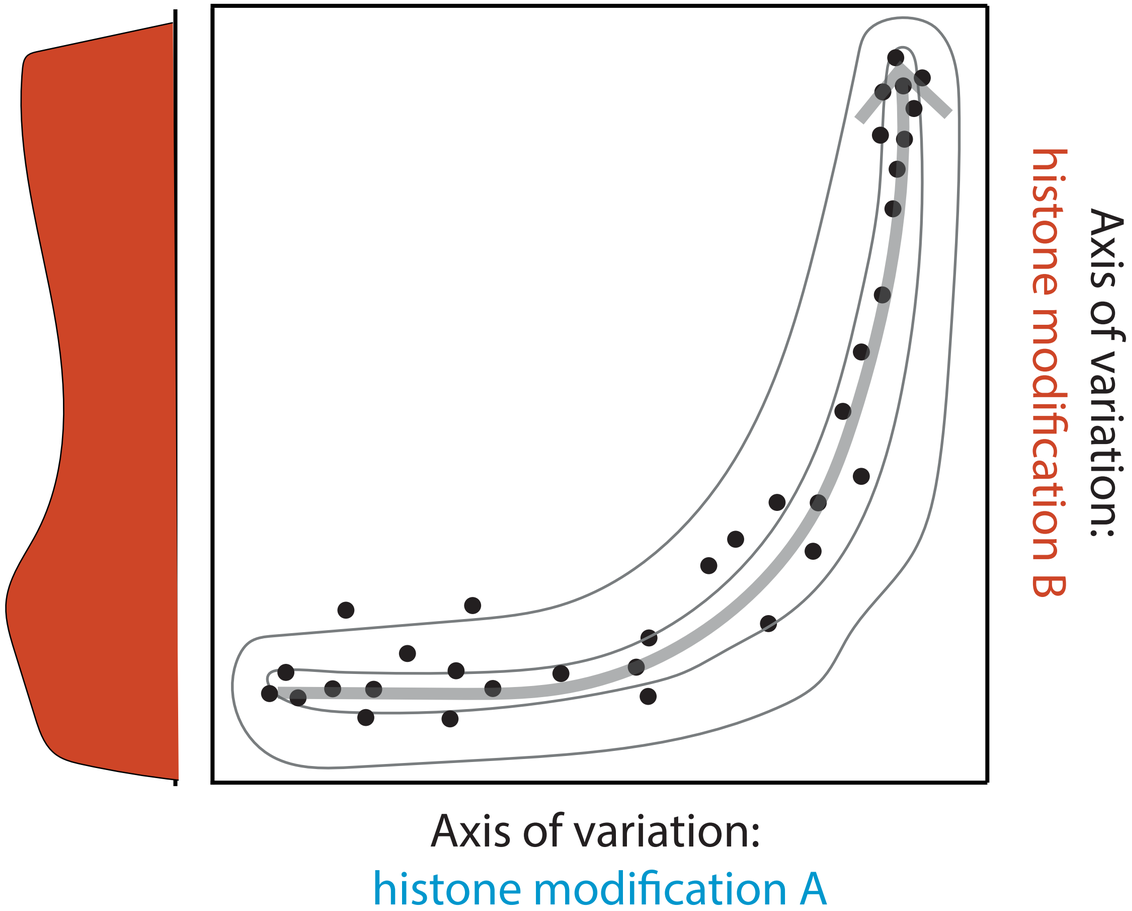
Axis of variation:
histone modification B

Axis of variation:
histone modification A

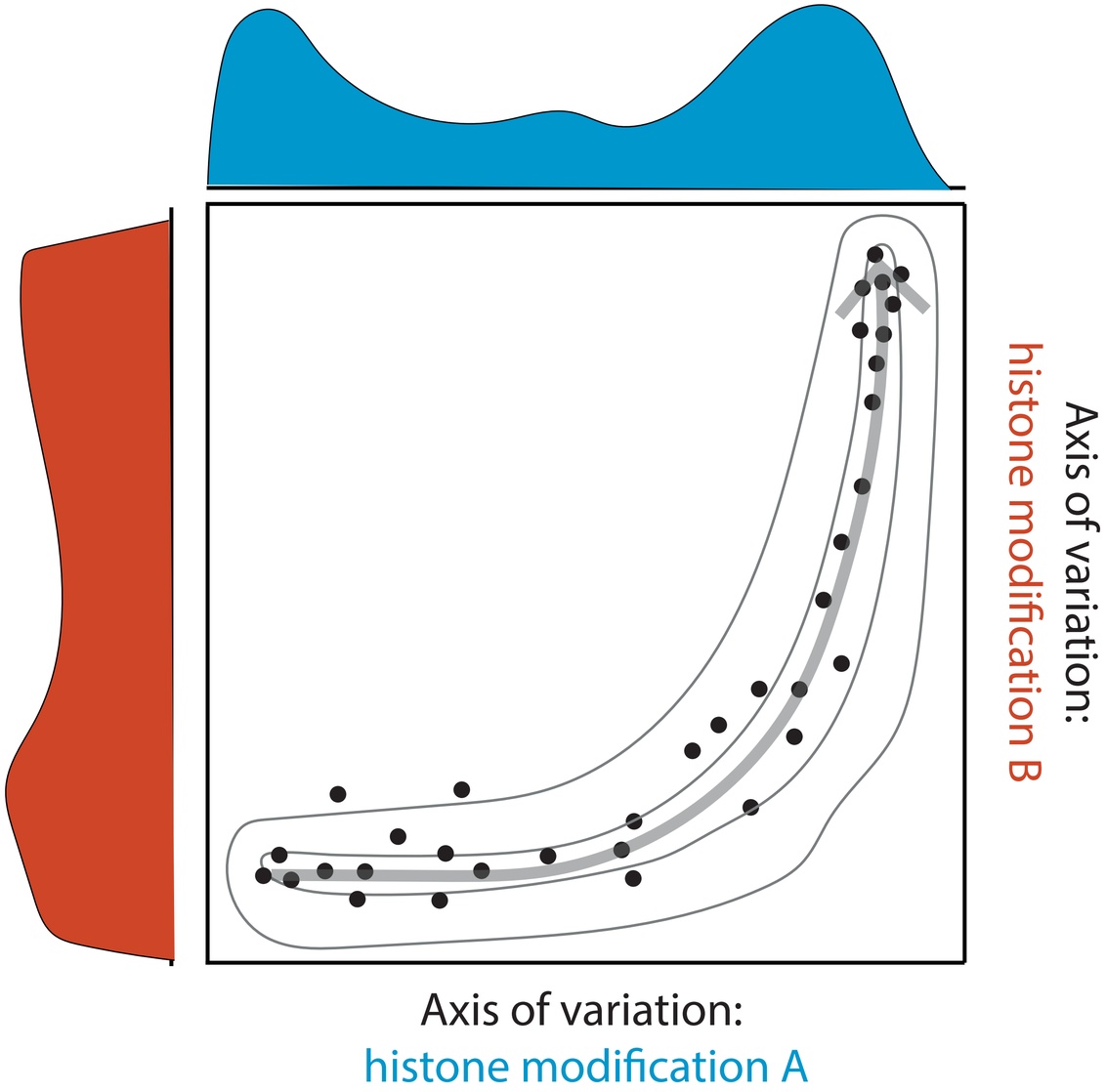
scChIX-seq can uncover relationships *between* histone modifications



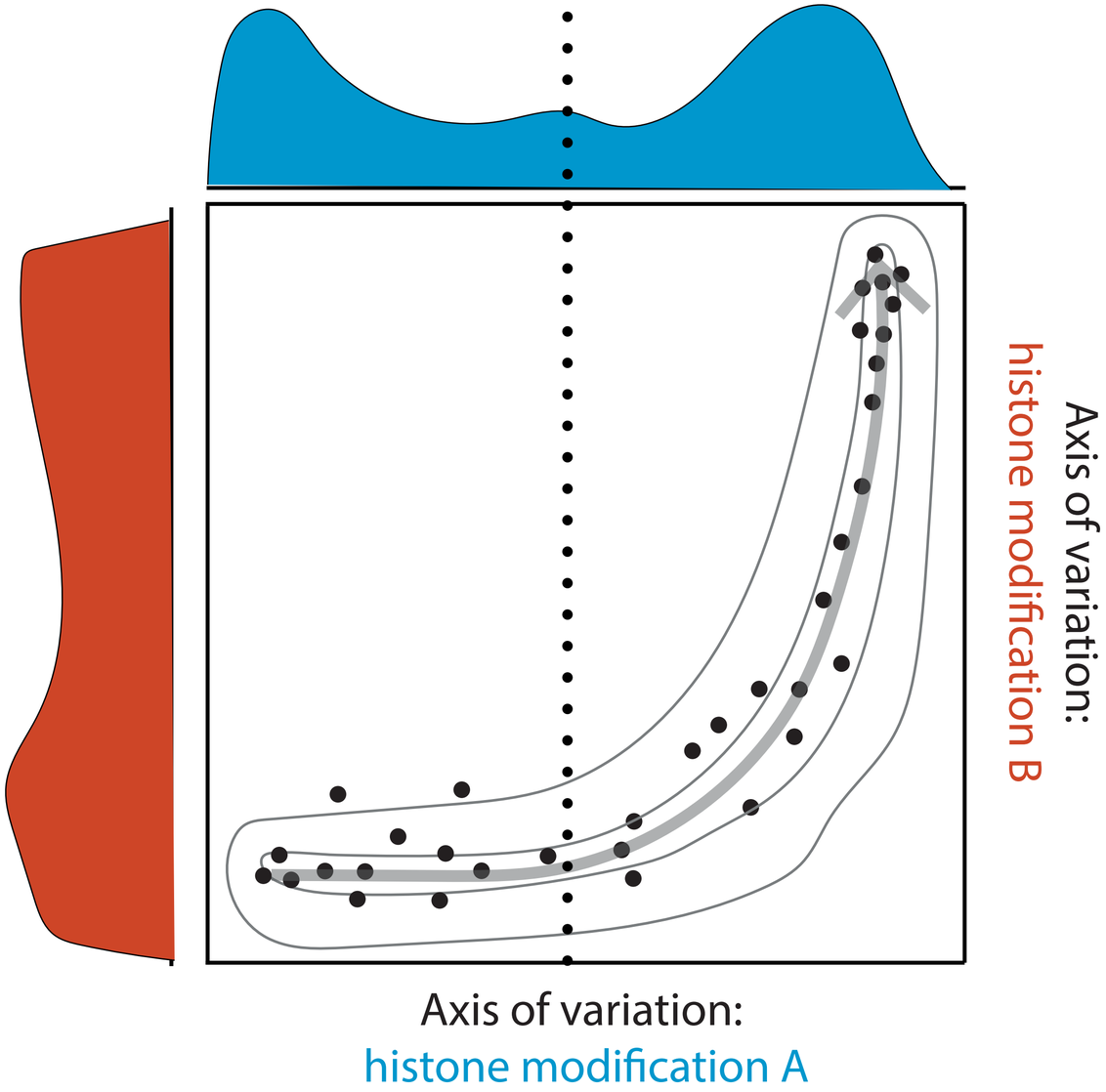
scChIX-seq can uncover relationships *between* histone modifications



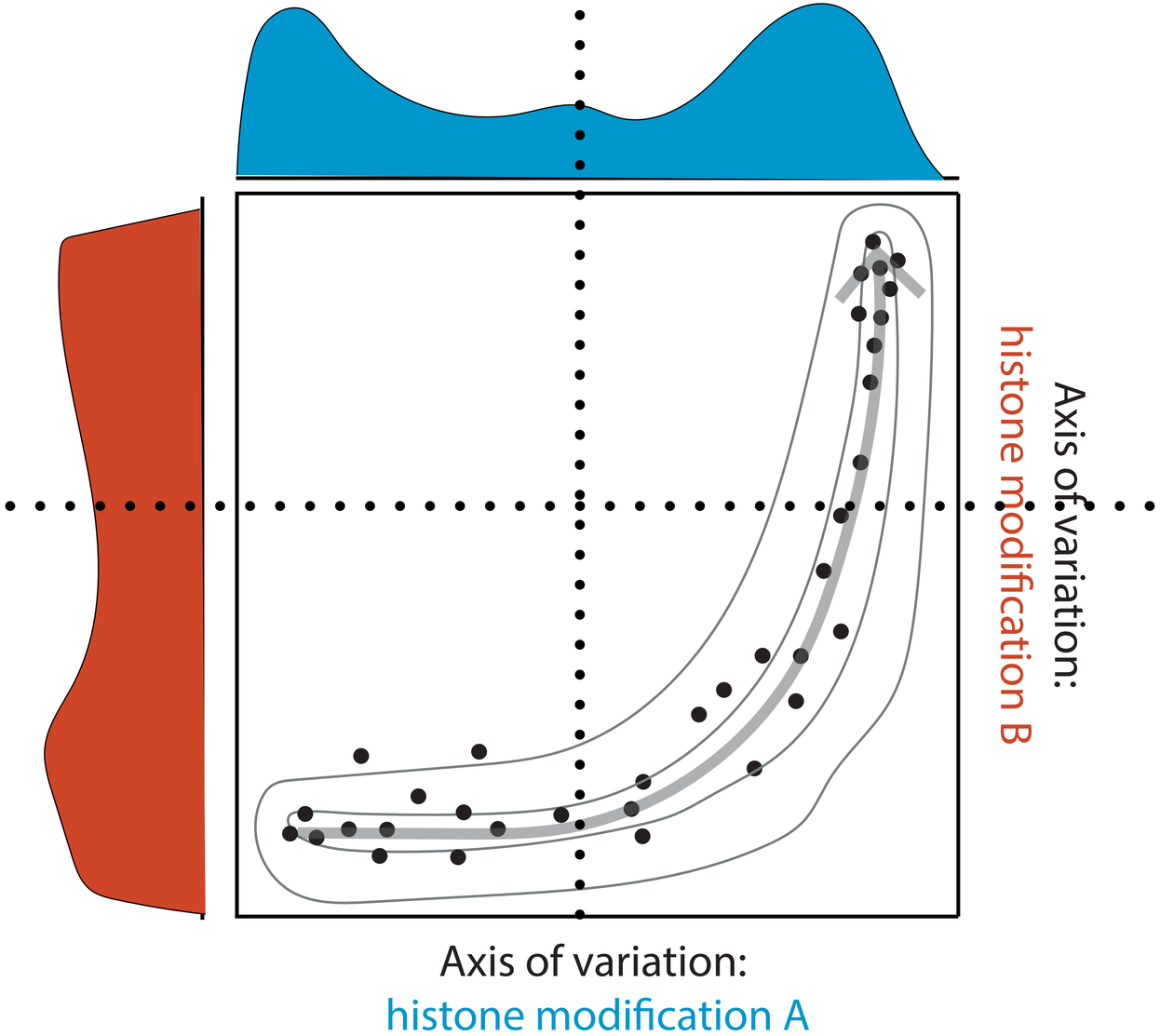
scChIX-seq can uncover relationships *between* histone modifications



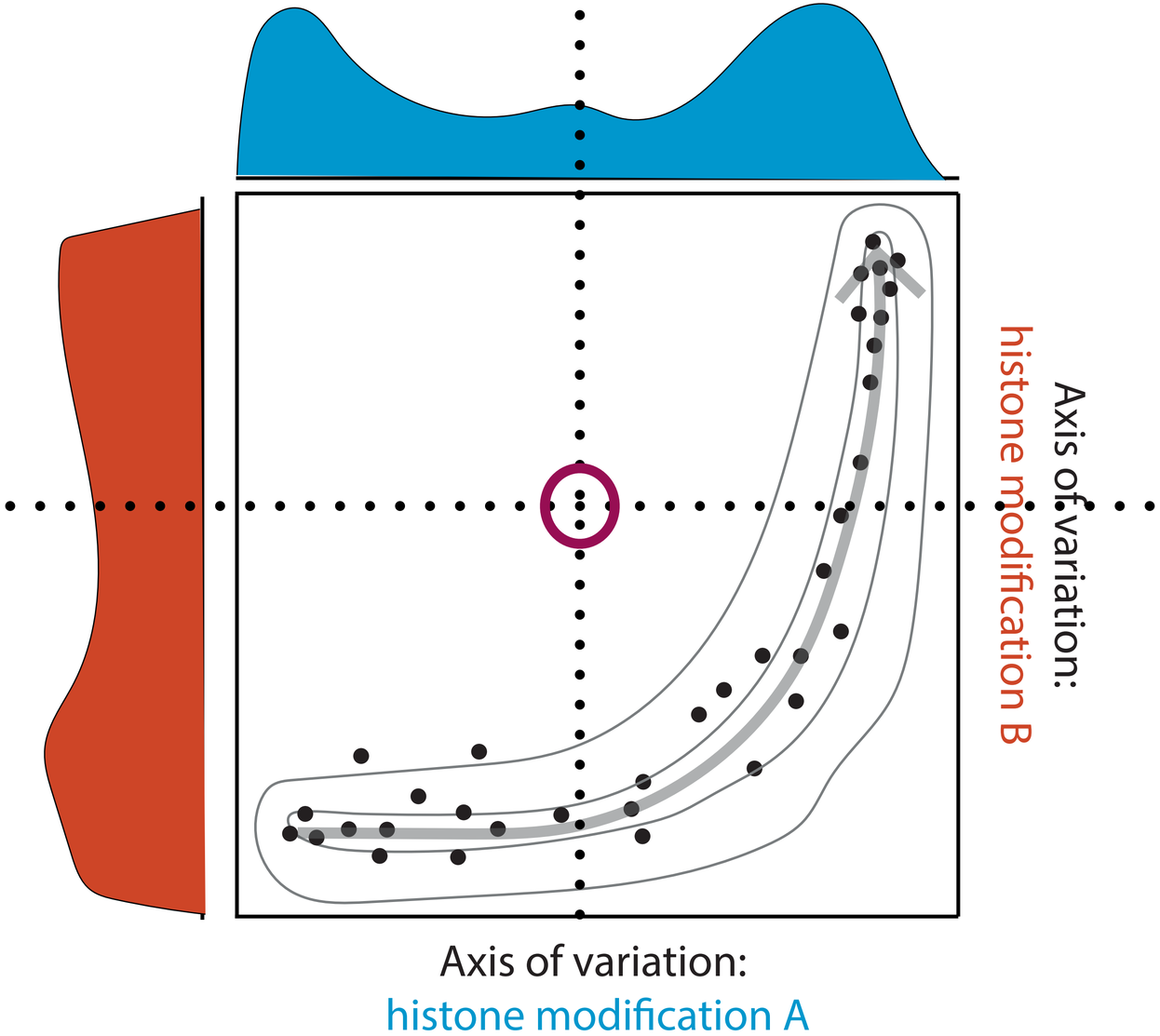
scChIX-seq can uncover relationships *between* histone modifications



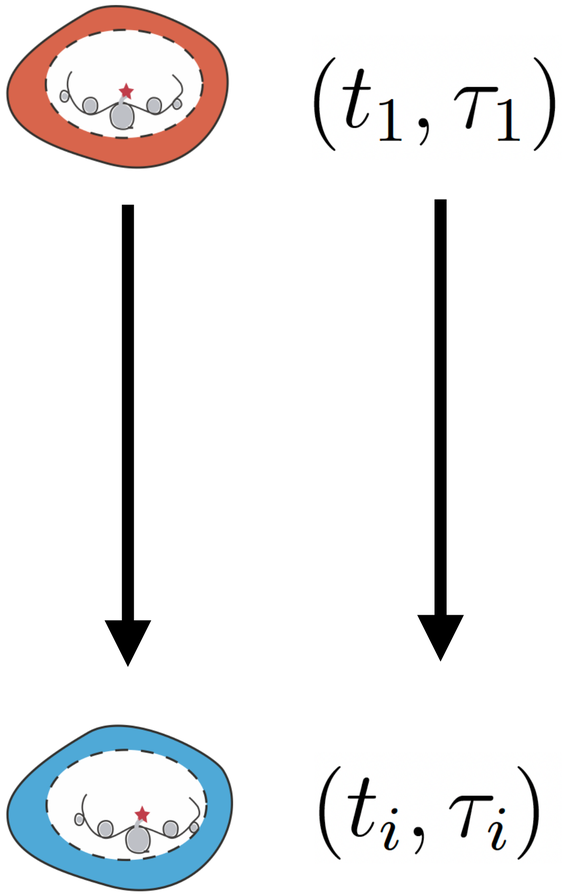
scChIX-seq can uncover relationships *between* histone modifications



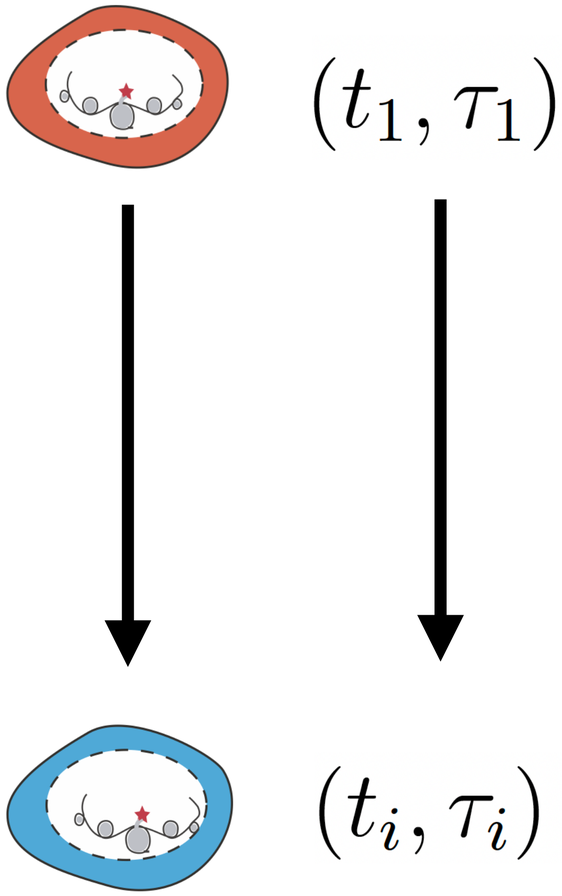
scChIX-seq can uncover relationships *between* histone modifications



Inferring pseudotime along both H3K4me1 and H3K36me3 reveals distinct dynamics



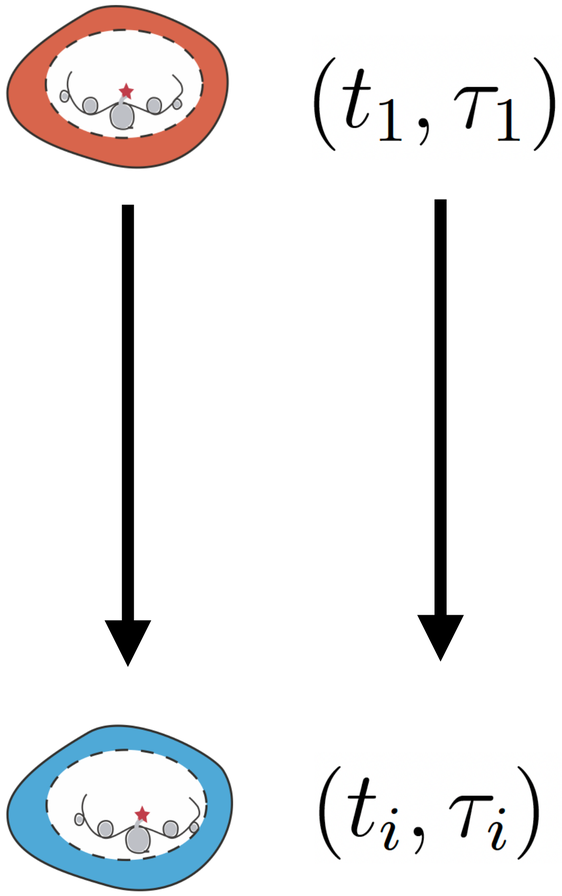
Inferring pseudotime along both H3K4me1 and H3K36me3 reveals distinct dynamics

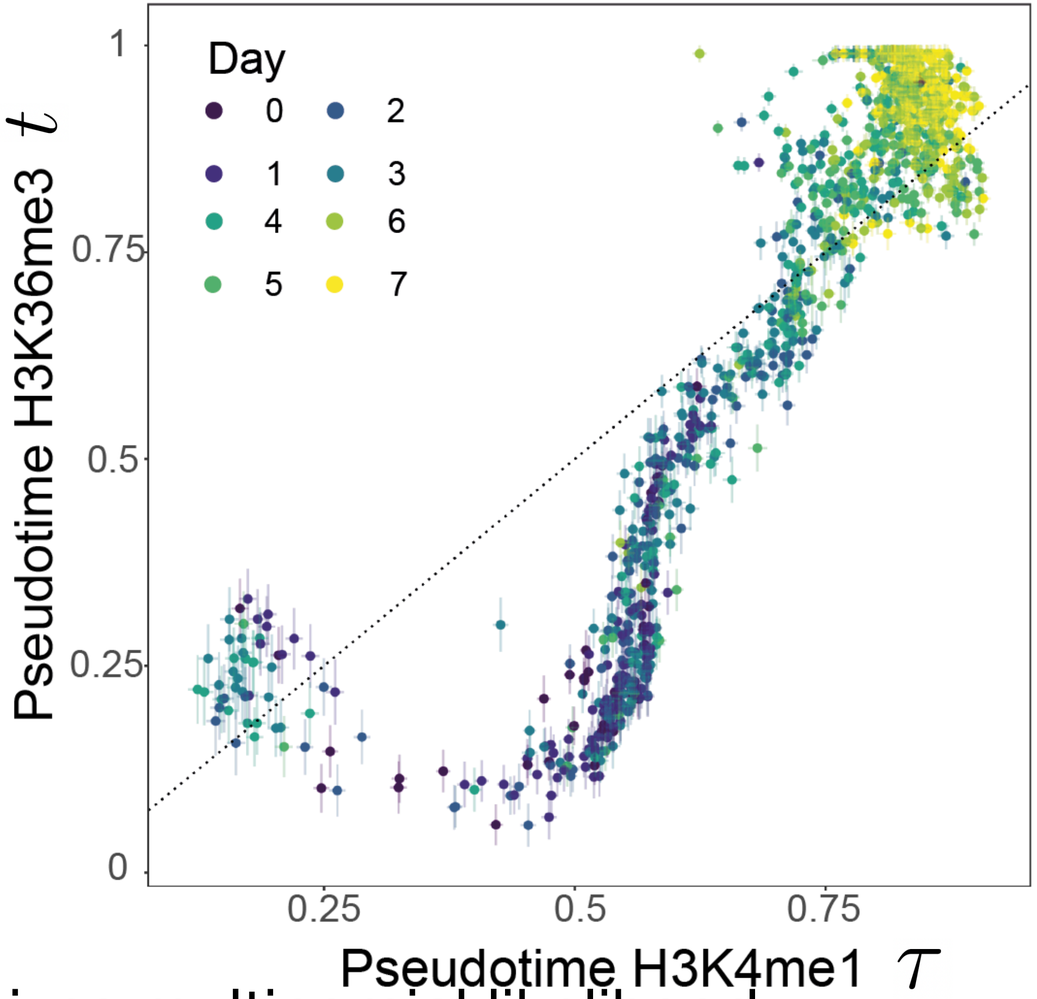


Find t and $\boldsymbol{\tau}$ that maximizes multinomial likelihood:

$$L(t, \tau) = \log\left(\Pr\left(\vec{y} | \vec{p}(t), \vec{q}(\tau), w\right)\right) \propto \sum_{g=1}^{G} y_g \log\left(w\vec{p}_g(t) + (1-w)\,\vec{q}_g(\tau)\right)$$

Inferring pseudotime along both H3K4me1 and H3K36me3 reveals distinct dynamics

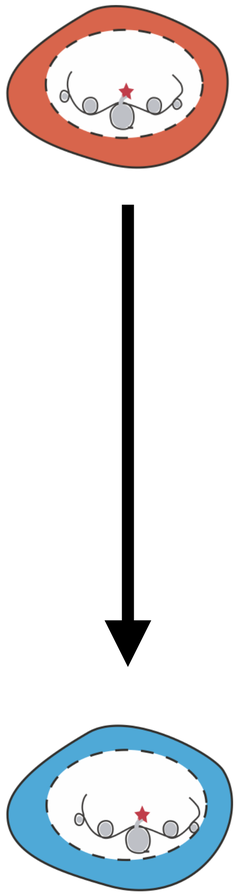




Find t and $\boldsymbol{\tau}$ that maximizes multinomial likelihood:

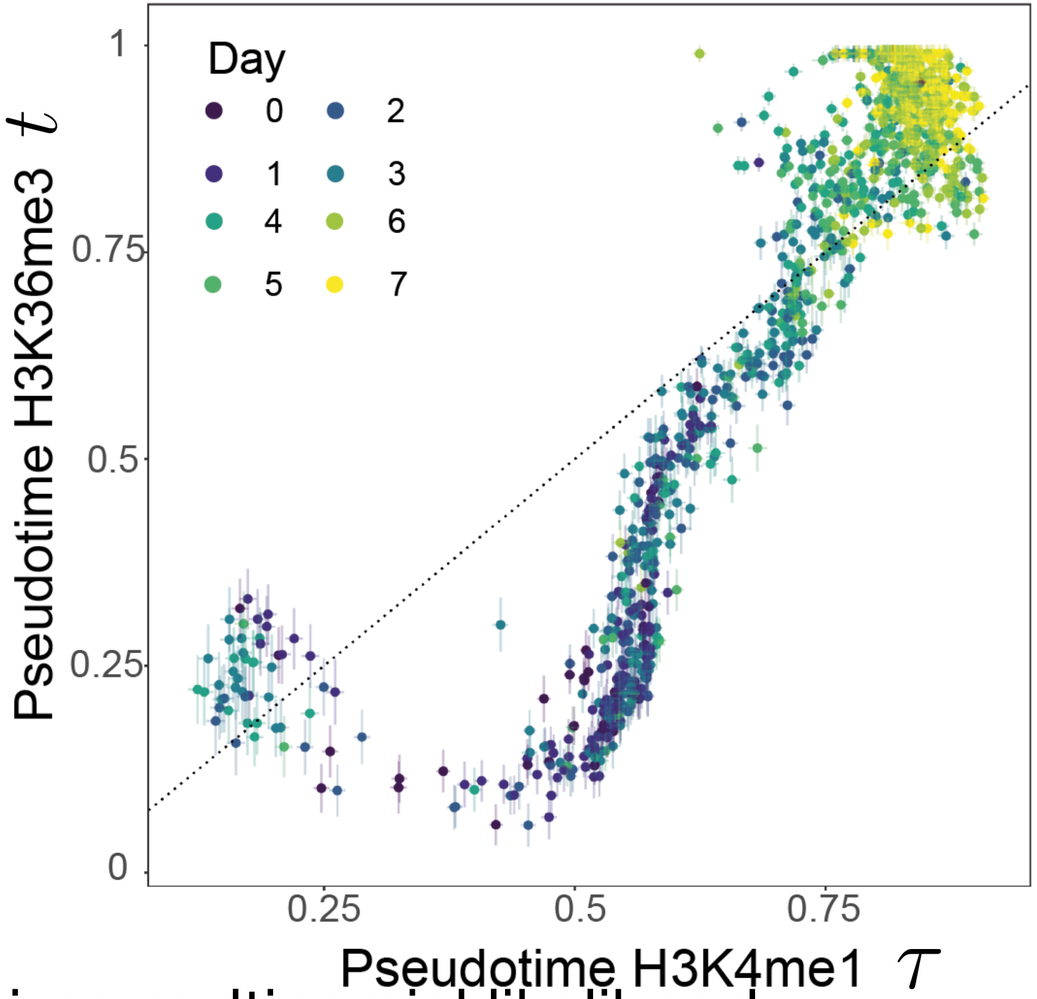
$$L(t,\tau) = \log\left(\Pr\left(\vec{y}|\vec{p}(t), \vec{q}(\tau), w\right)\right) \propto \sum_{g=1}^{G} y_g \log\left(w\vec{p}_g(t) + (1-w)\,\vec{q}_g(\tau)\right)$$

Inferring pseudotime along both H3K4me1 and H3K36me3 reveals distinct dynamics



(t_1, τ_1)

(t_i, τ_i)



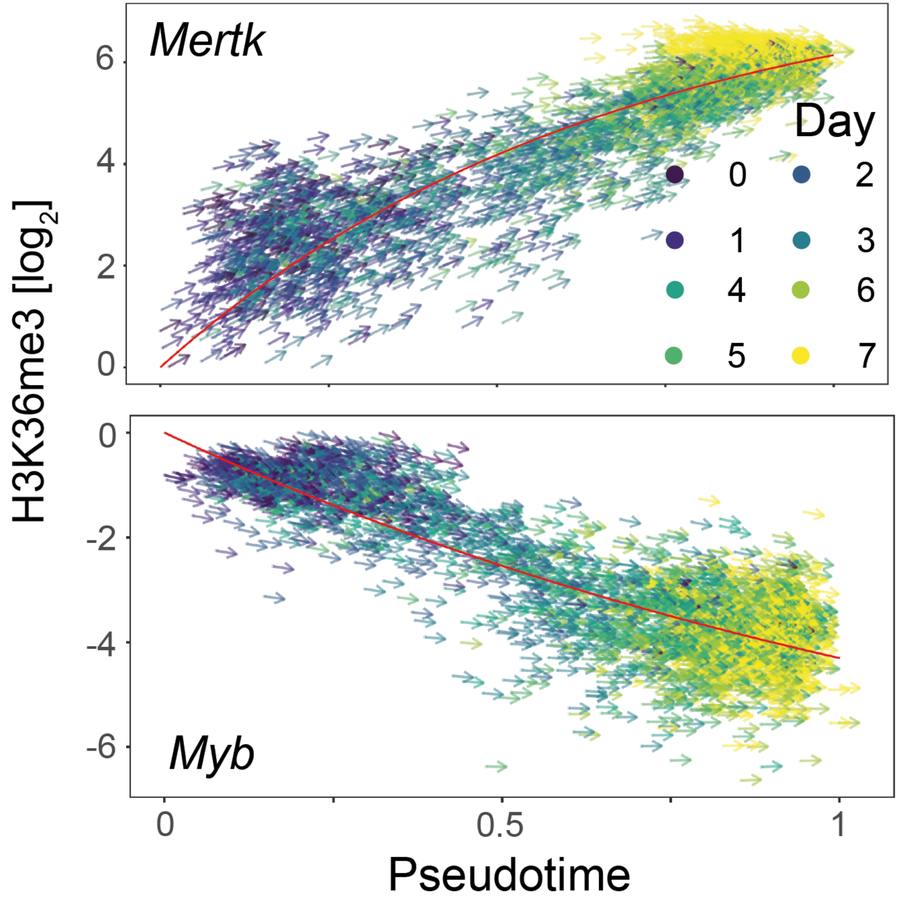
K4me1 primes genes for transcription (K36me3)

Find t and $\boldsymbol{\tau}$ that maximizes multinomial likelihood:

$$L(t,\tau) = \log\left(\Pr\left(\vec{y}|\vec{p}(t), \vec{q}(\tau), w\right)\right) \propto \sum_{g=1}^{G} y_g \log\left(w\vec{p}_g(t) + (1-w)\,\vec{q}_g(\tau)\right)$$

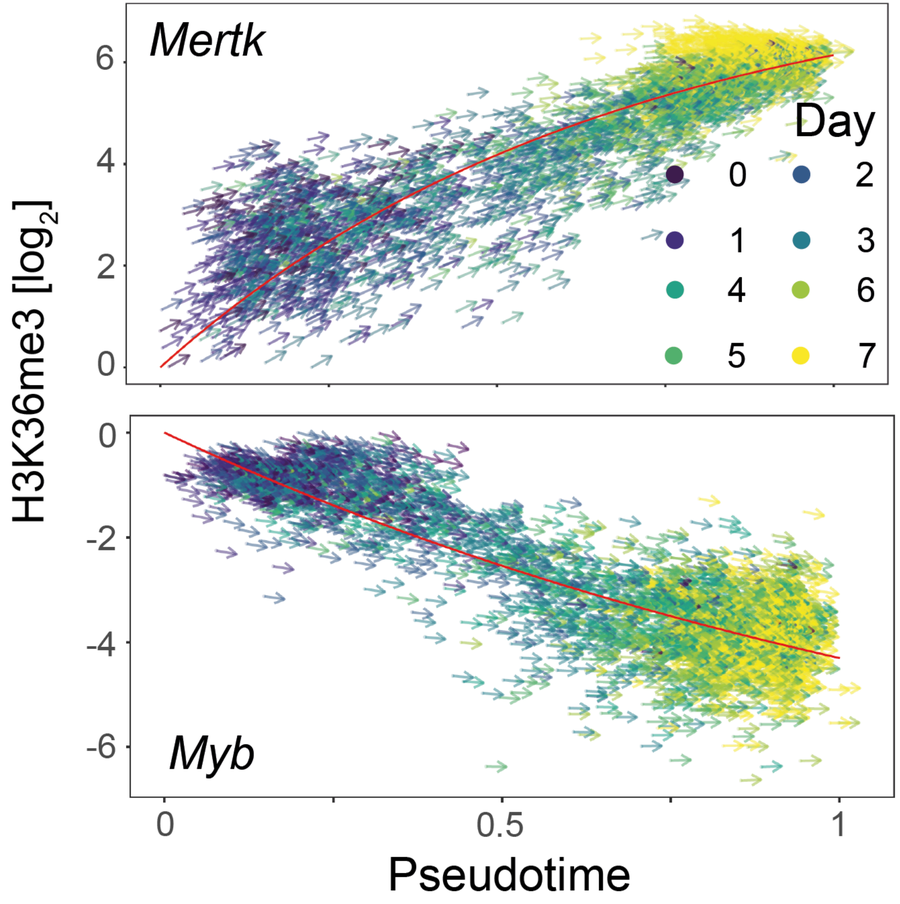
Modeling the dynamics of both histone modifications reveals chromatin velocity

$$\frac{dK_{36}(t)}{dt} = K_4(t) - \gamma K_{36}(t)$$

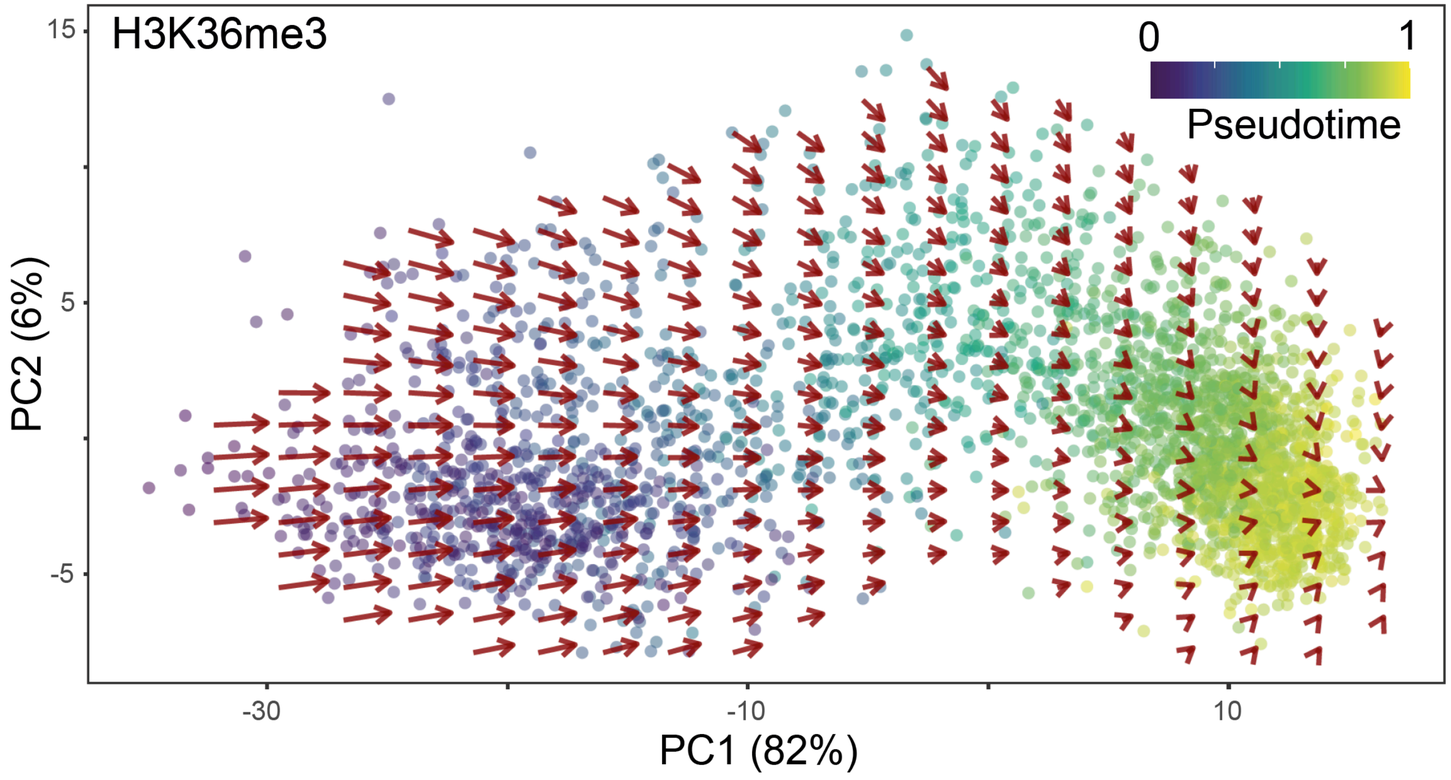


Modeling the dynamics of both histone modifications reveals chromatin velocity

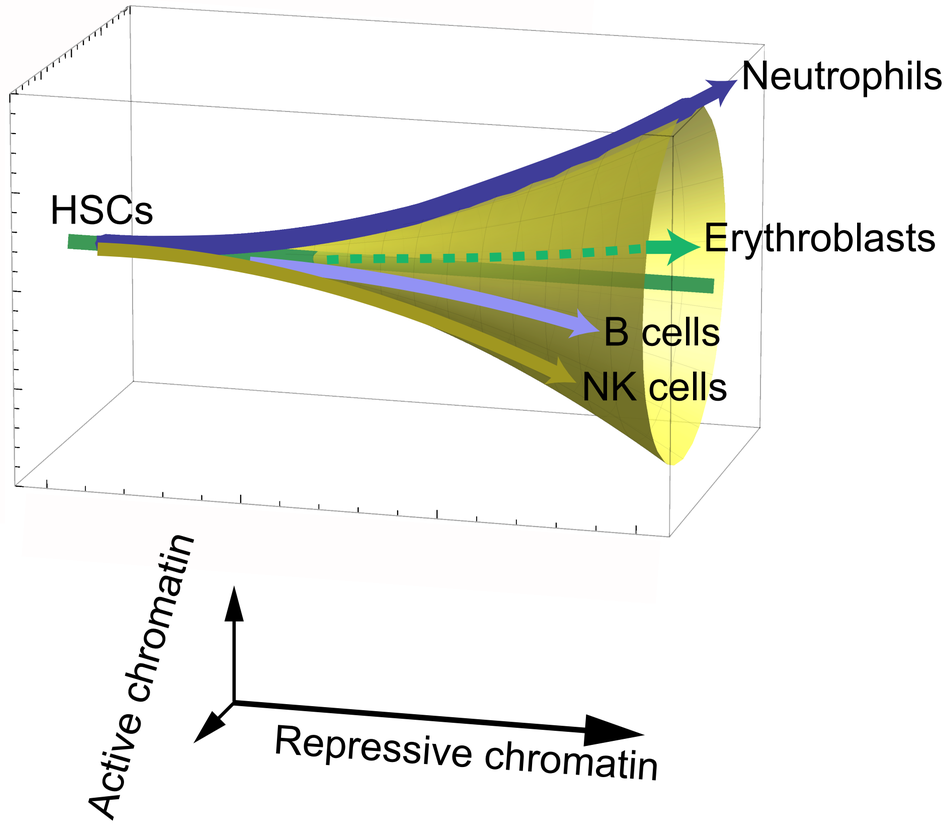
$$\frac{dK_{36}(t)}{dt} = K_4(t) - \gamma K_{36}(t)$$

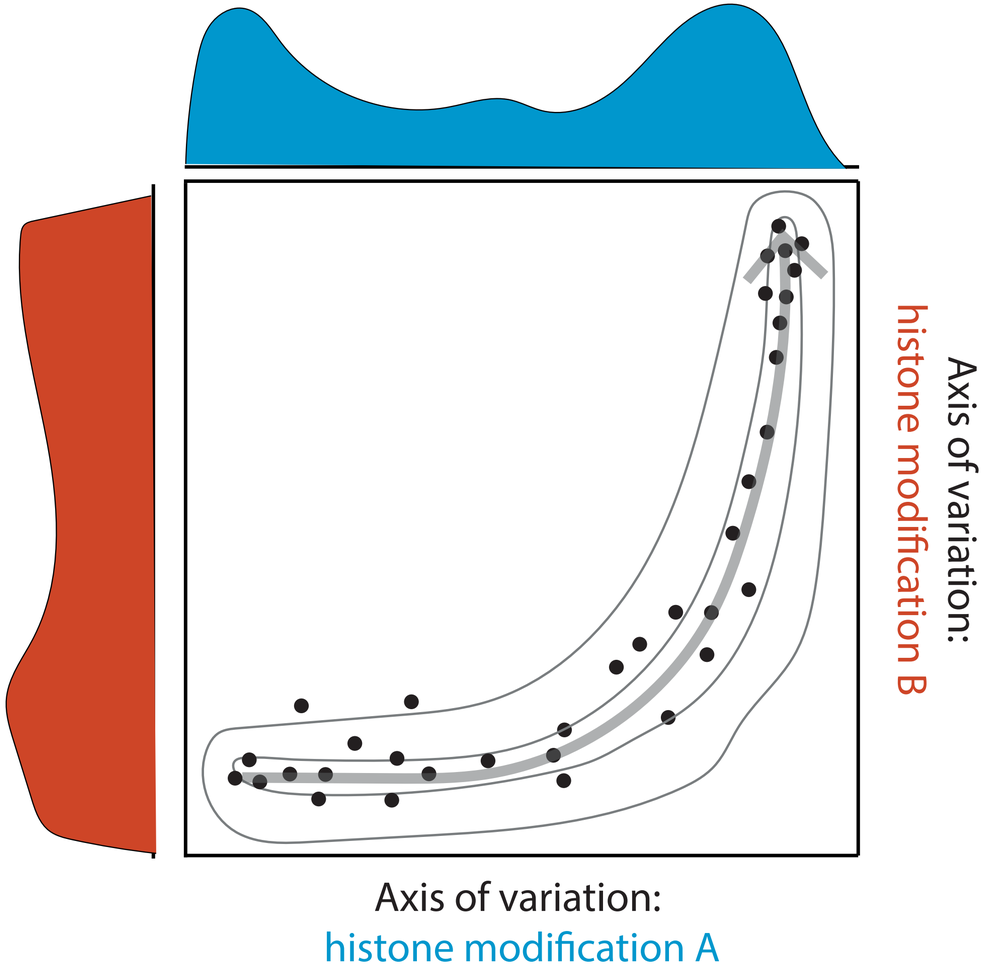


Summary of 206 genes



Integrative methods reveal interactions that are "greater than the sum of the parts"





sortChIC: Zeller*, Yeung* et al. *Nat Genetics* 2022

scChIX-seq: Yeung*, Florescu*, et al. *Nat Biotech* 2023

Challenges: towards data science-driven experimental methods and design

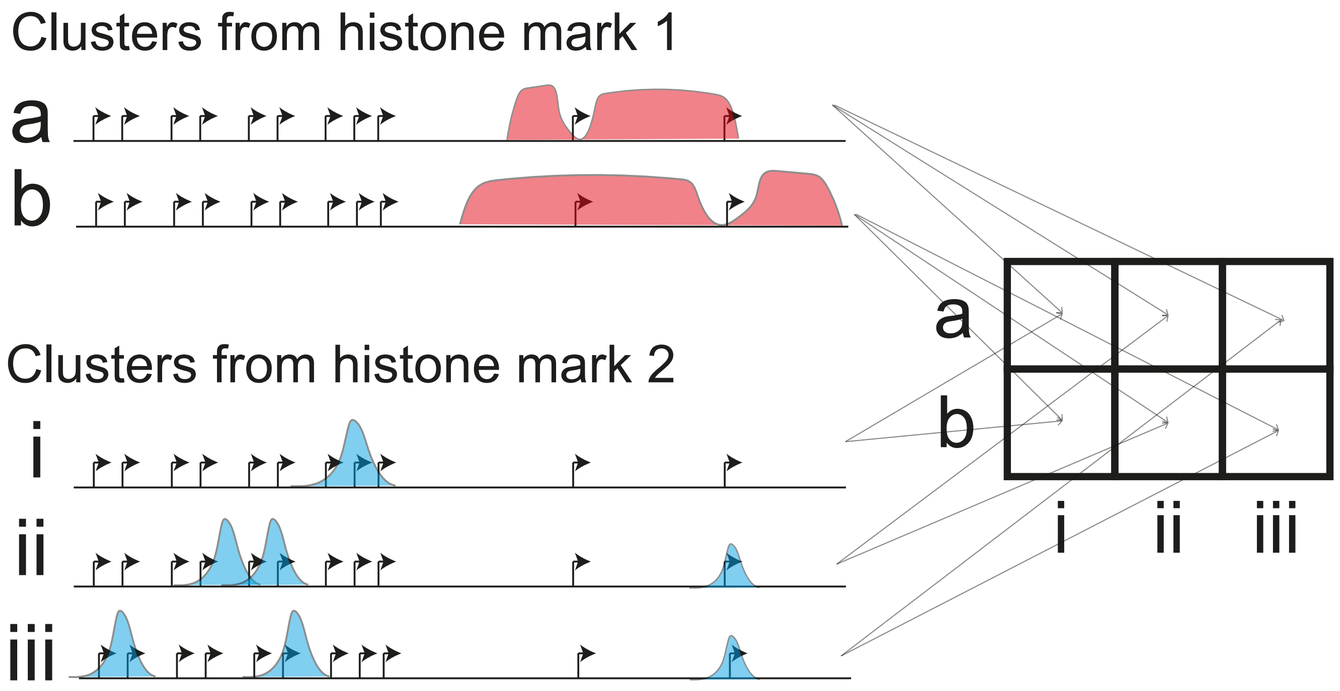
- Data science-driven solutions can reveal experimental insights that expand the gene regulatory picture captured by single-cell genomics.

- Dynamics of different chromatin states can be distinct: why and how much they differ influences experimental design and integrative analysis.

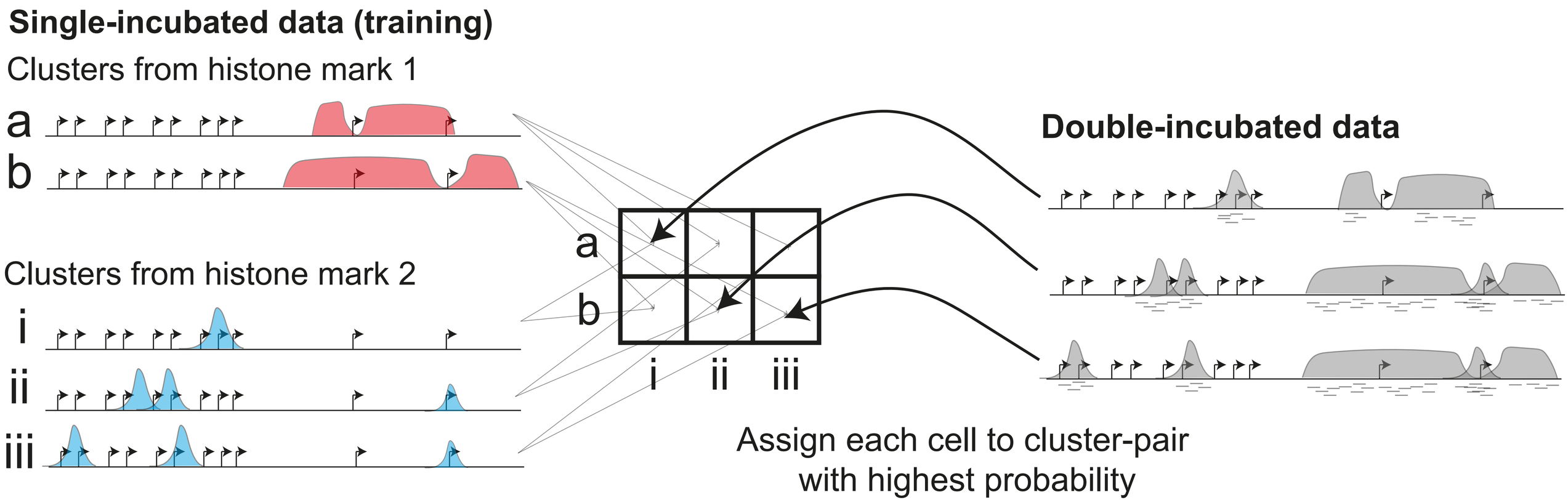
- What are the limits of analyzing noisy snapshot data to learn the real underlying stochastic trajectories? Can they be (partially) alleviated?

We use single-incubated data as training to infer cell type and heterochromatin identity in double-incubated cells

Single-incubated data (training)

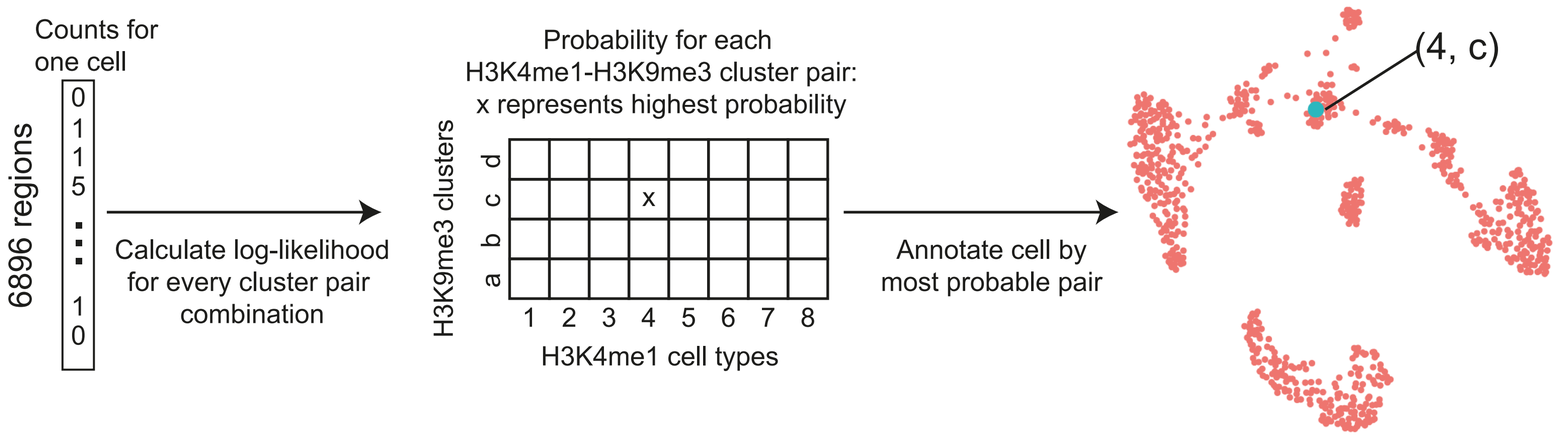


We use single-incubated data as training to infer cell type and heterochromatin identity in double-incubated cells



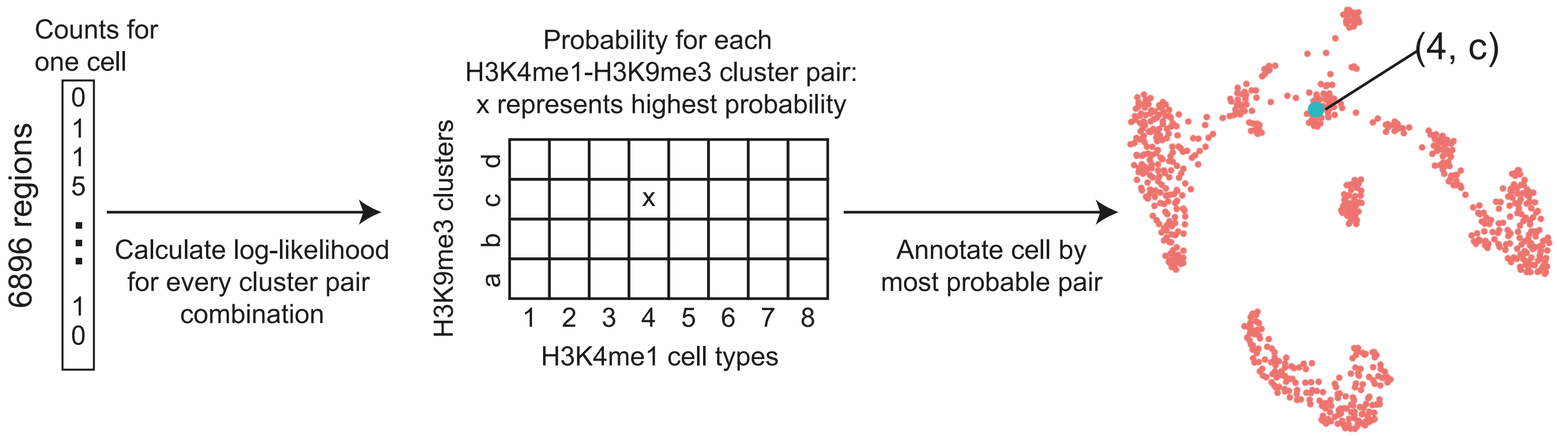
Each double-incubated cell generates a likelihood grid, which gives probabilities for each cluster-pair

Double-incubated single cells (observed)

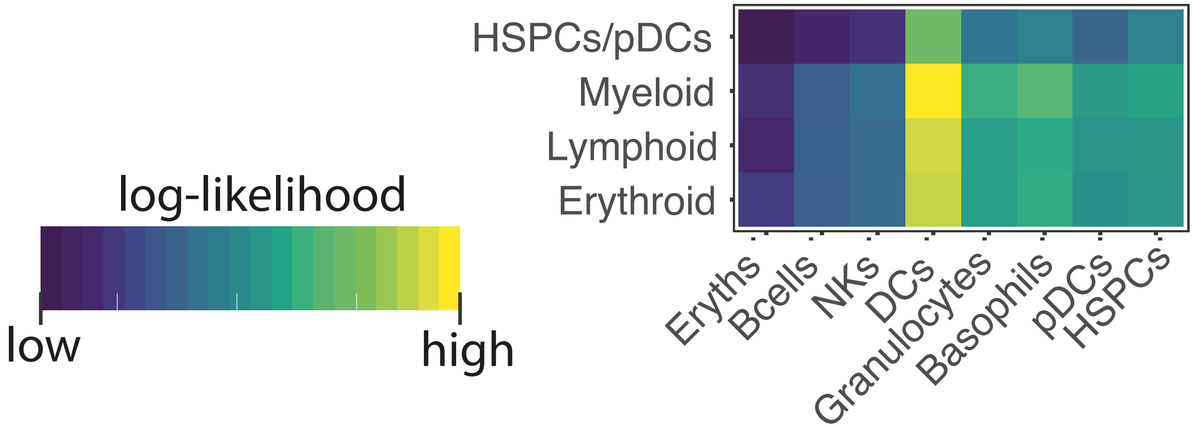


Each double-incubated cell generates a likelihood grid, which gives probabilities for each cluster-pair

Double-incubated single cells (observed)

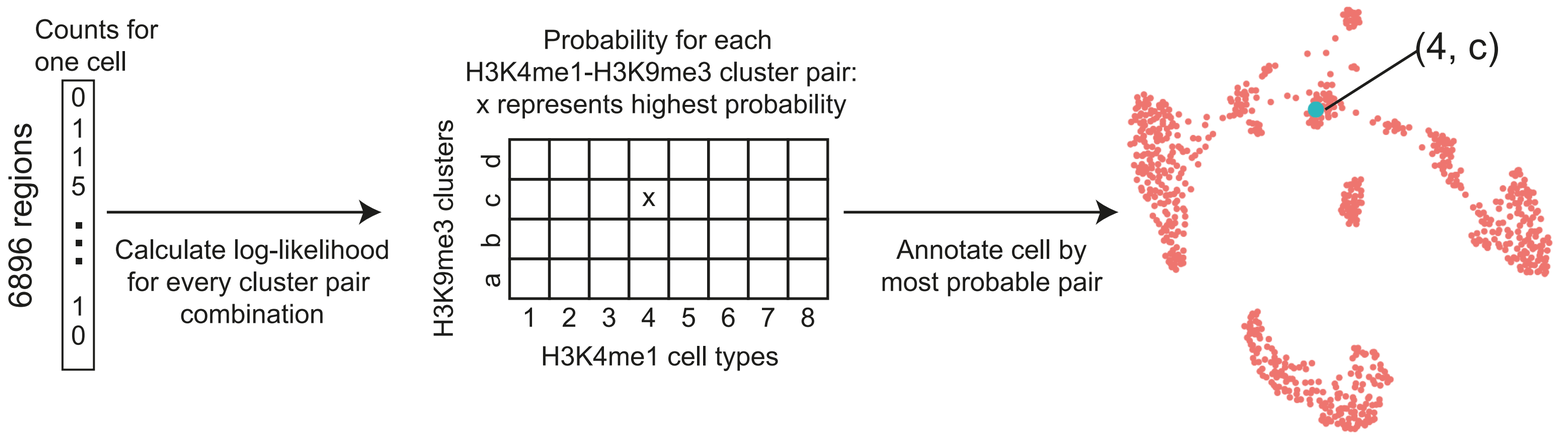


Likelihood map for one cell

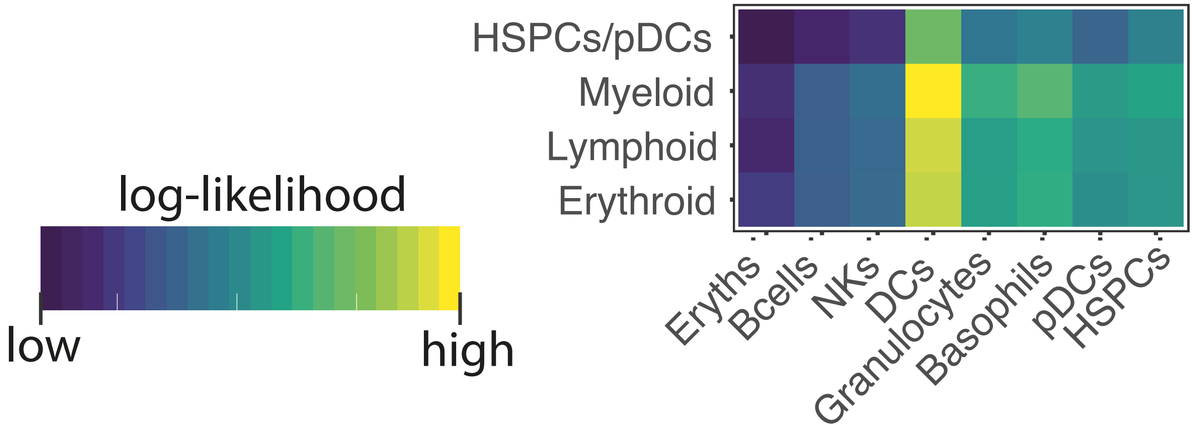


Each double-incubated cell generates a likelihood grid, which gives probabilities for each cluster-pair

Double-incubated single cells (observed)



Likelihood map for one cell



Example: likelihood at (4, c) is:

$$\text{logLikelihood}|\vec{p}_4, \vec{p}_c \propto \sum_{k=1}^{K} y_k \log\left(w p_{4,k} + (1-w)\, p_{c,k}\right)$$

K9 probability at region *k*

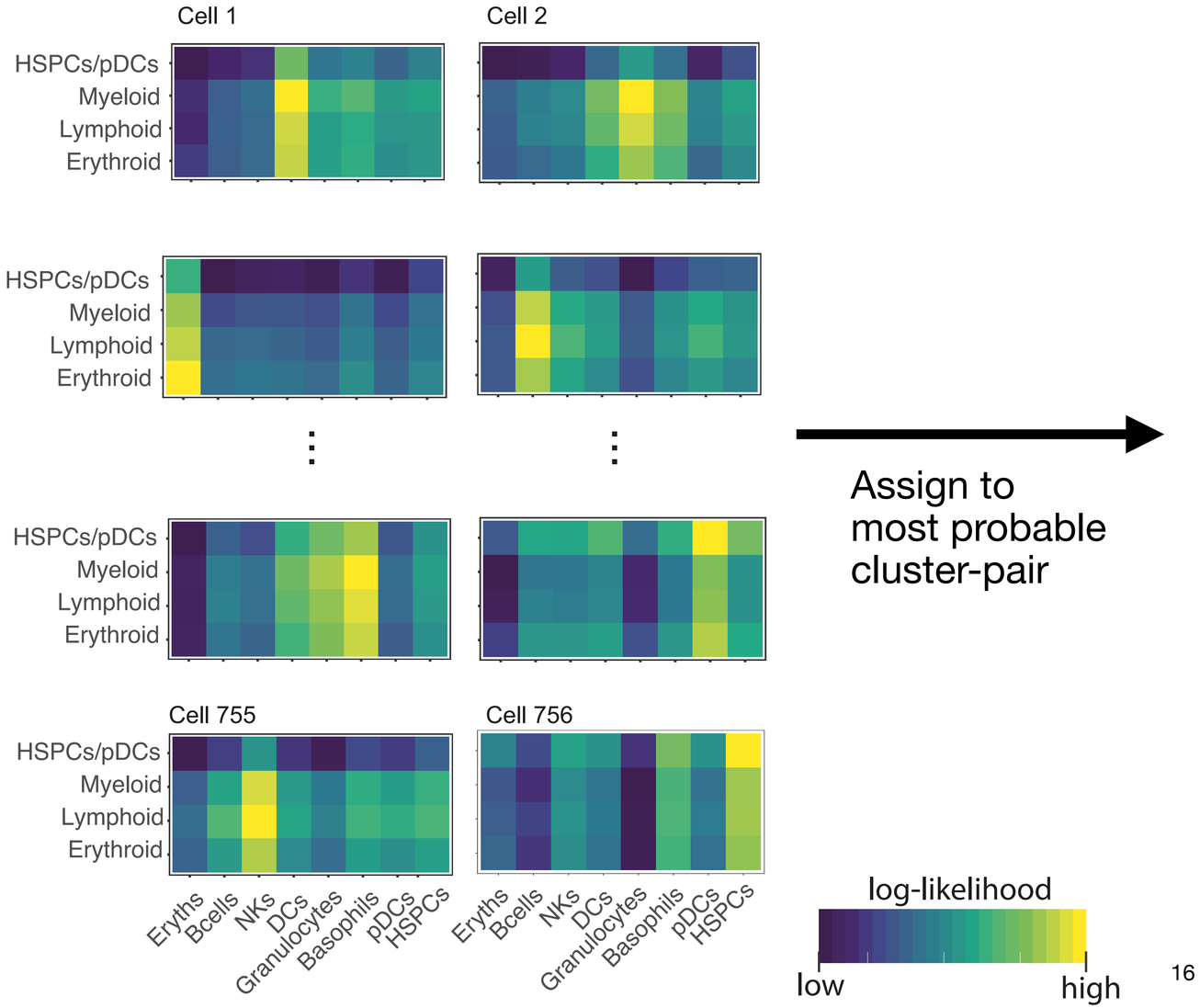
K4 probability at region *k*

Mixing weight (inferred)

Double cuts at region *k*

Double-incubated analysis reveals heterochromatin can be shared across related cell types

Calculate logLikelihood grid for each double-incubated cell:

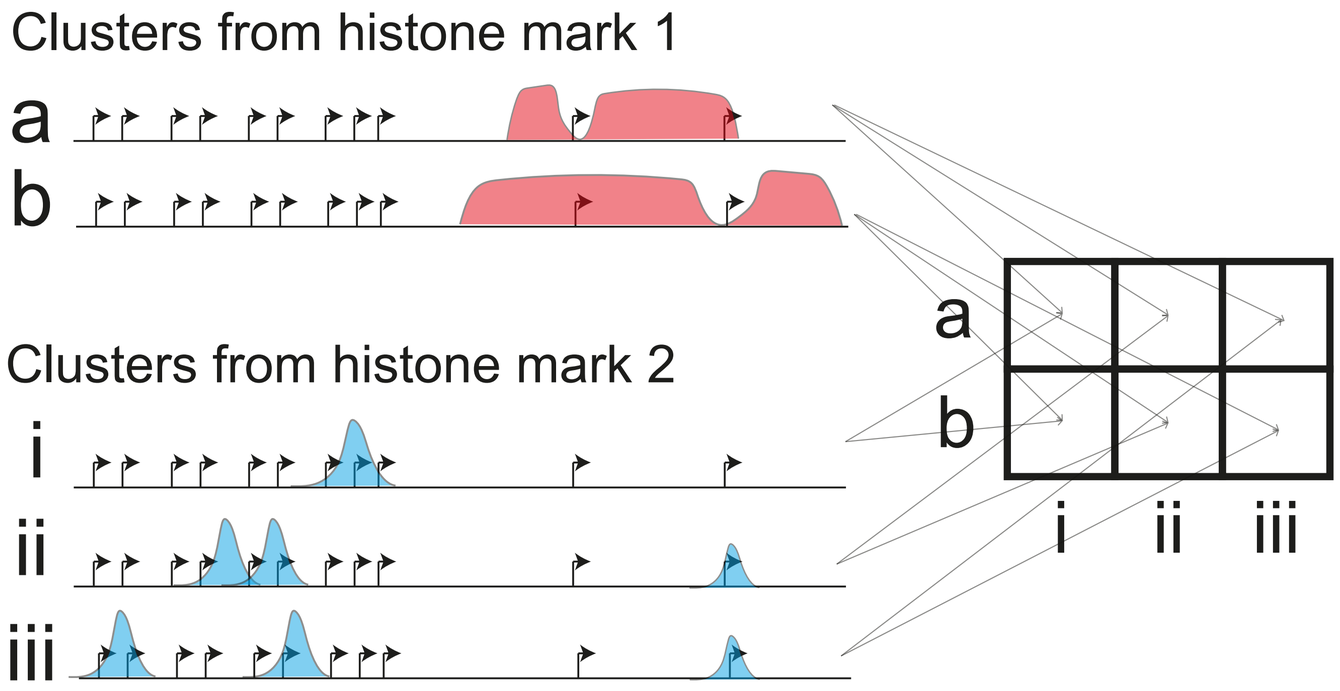


Double-incubated analysis reveals heterochromatin can be shared across related cell types

Calculate logLikelihood grid for each double-incubated cell:

We use single-incubated data as training to infer cell type and heterochromatin in double-incubated cells

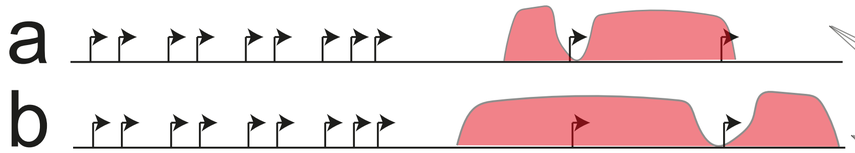
Single-incubated data (training)



We use single-incubated data as training to infer cell type and heterochromatin in double-incubated cells

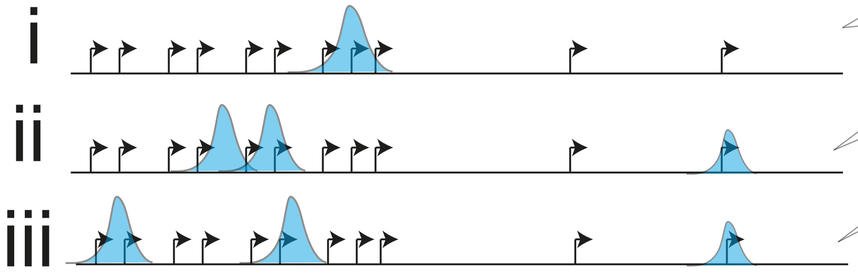
Single-incubated data (training)

Clusters from histone mark 1

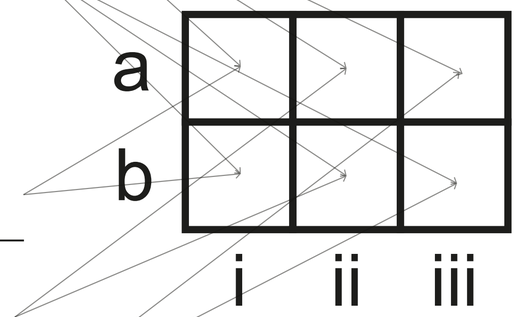


$\vec{p}_a = [0.01, 0.05, ..., 0.3]$

Clusters from histone mark 2



$\vec{p}_{ii} = [0.3, 0.05, ..., 0.01]$

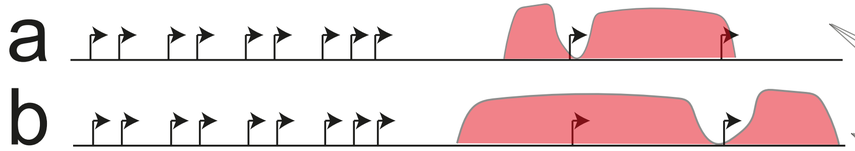


LDA gives these probabilities for free

We use single-incubated data as training to infer cell type and heterochromatin in double-incubated cells

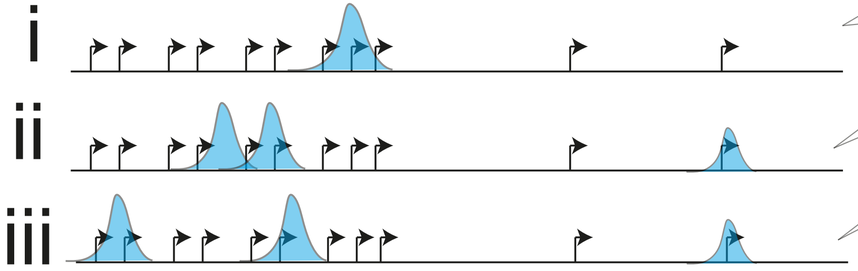
Single-incubated data (training)

Clusters from histone mark 1

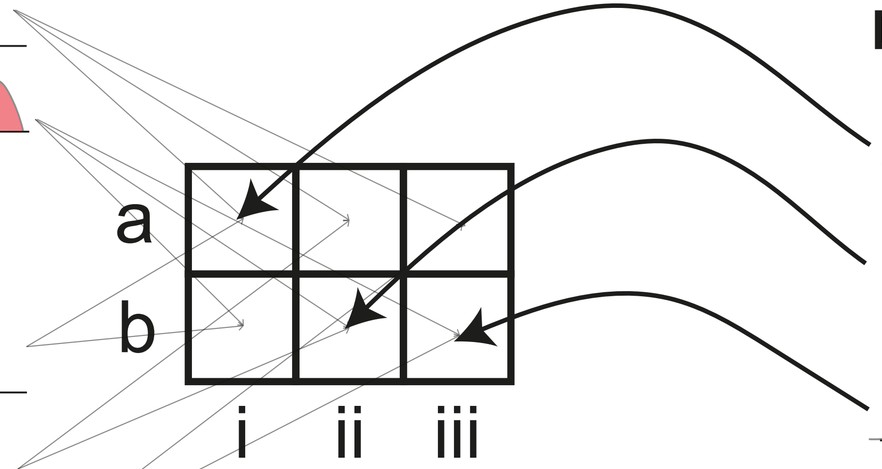


$\vec{p}_a = [0.01, 0.05, ..., 0.3]$

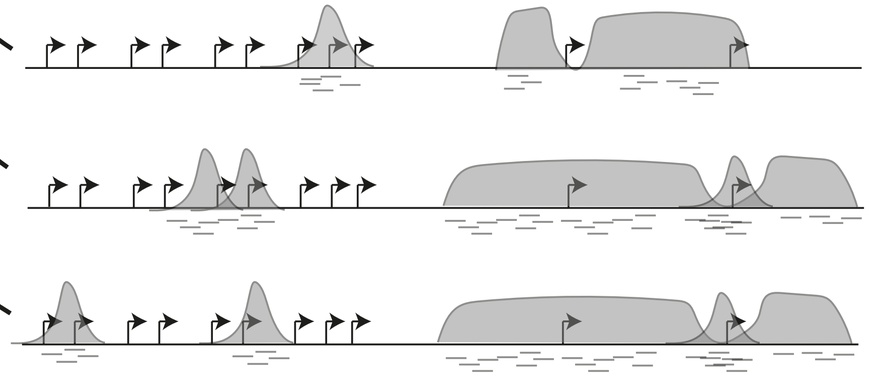
Clusters from histone mark 2



$\vec{p}_{ii} = [0.3, 0.05, ..., 0.01]$



Double-incubated data



Assign each cell to cluster-pair with highest probability

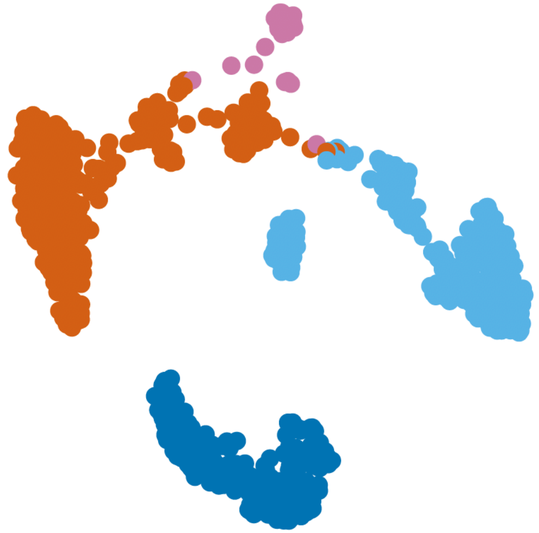
LDA gives these probabilities for free

Model for double-incubated counts coming from cluster b and ii:

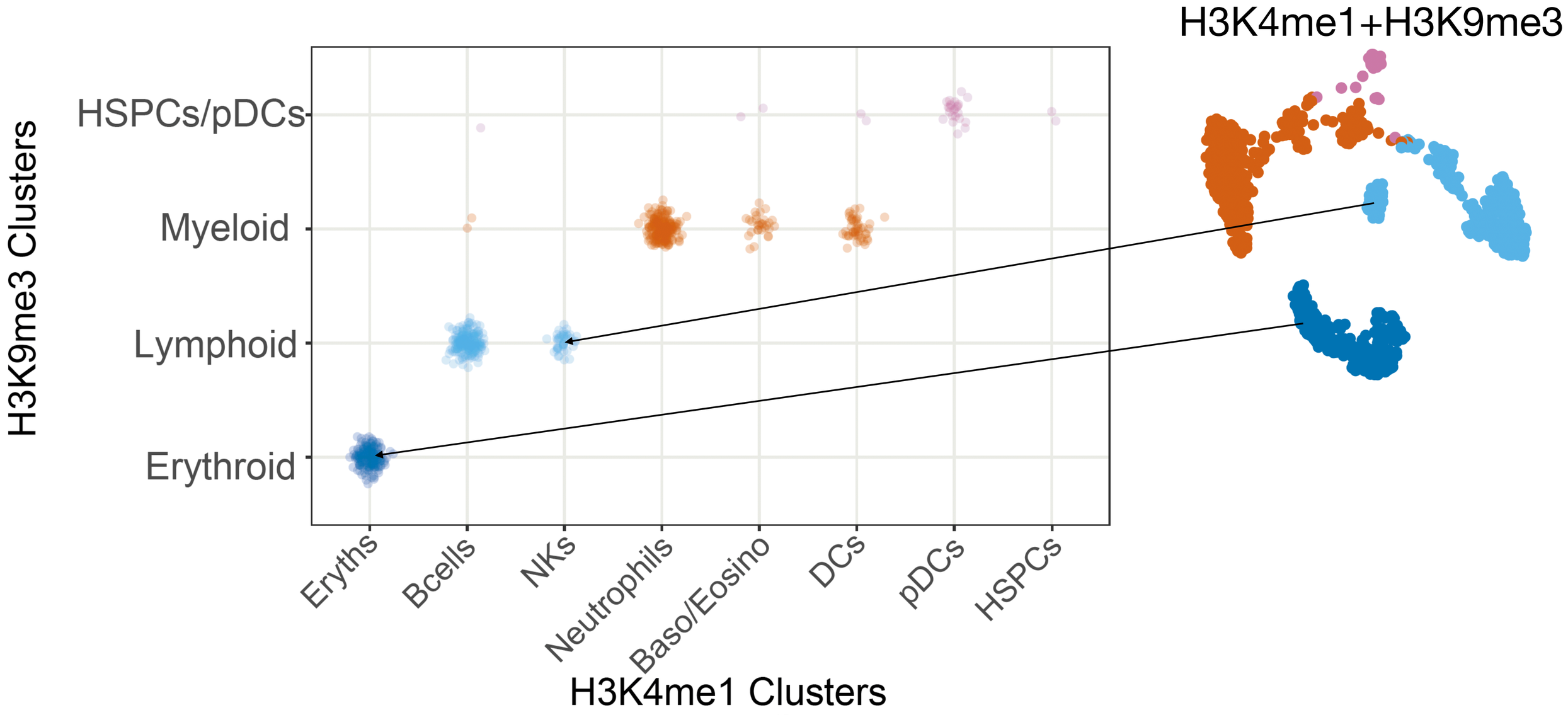
$$\vec{y}|\vec{p}_b, \vec{p}_{ii} \sim \text{Multinomial}\left(w\vec{p}_b + (1-w)\vec{p}_{ii}, N\right)$$

Distinct cell types from related lineage share similar heterochromatin

H3K4me1+H3K9me3



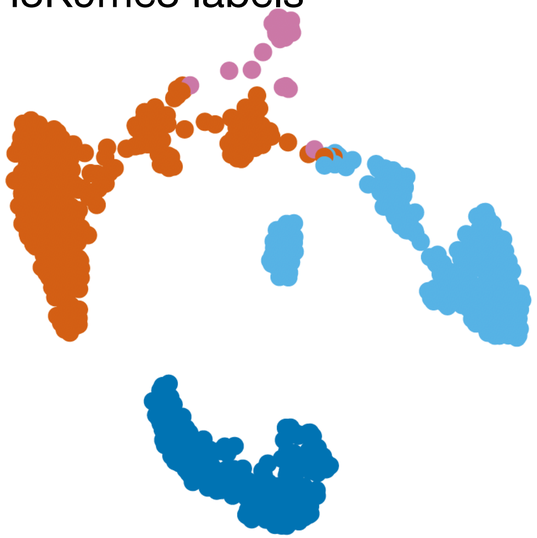
Distinct cell types from related lineage share similar heterochromatin



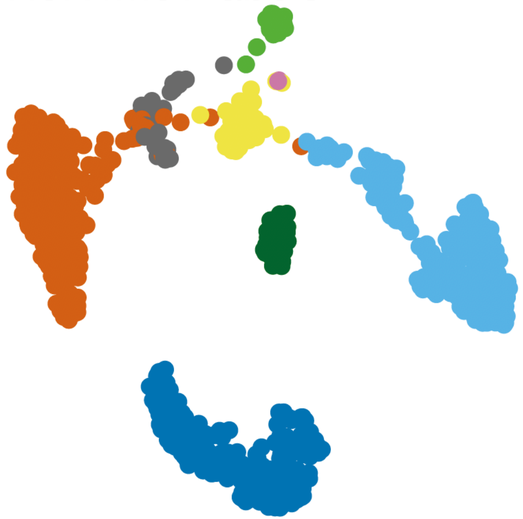
Chromatin regulation gives information of its cell type *and* its lineage

Each cell has two labels:

H3K9me3 labels



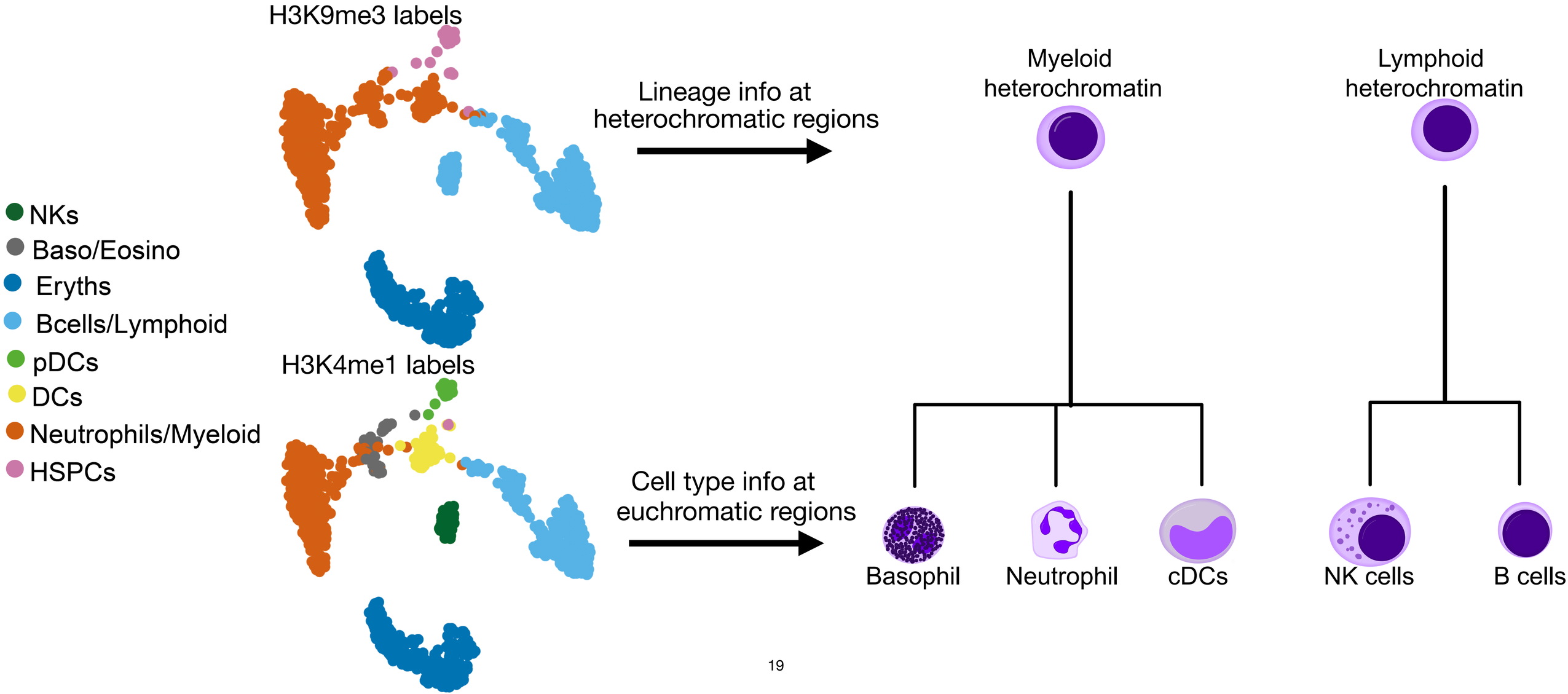
H3K4me1 labels



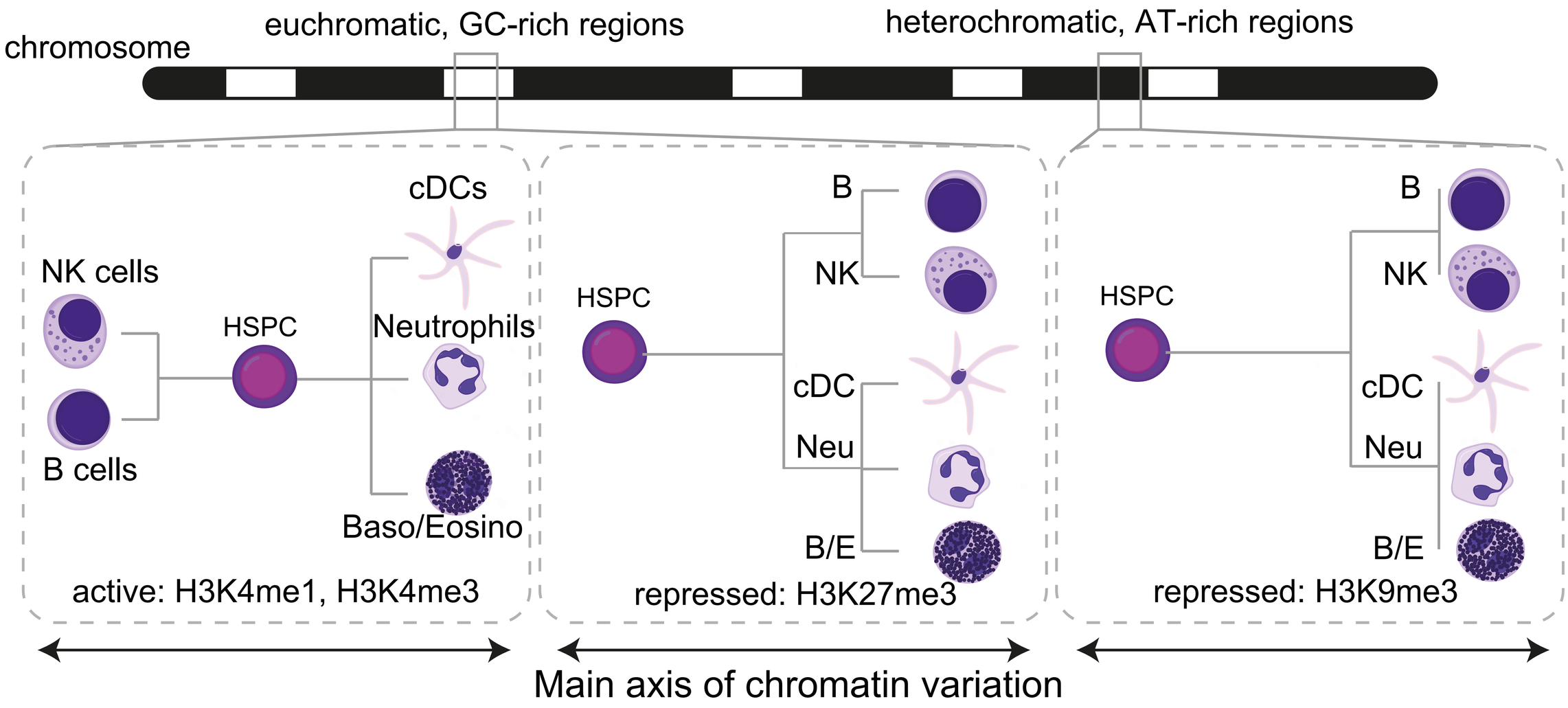
- NKs
- Baso/Eosino
- Eryths
- Bcells/Lymphoid
- pDCs
- DCs
- Neutrophils/Myeloid
- HSPCs

Chromatin regulation gives information of its cell type *and* its lineage

Each cell has two labels:



Repressive chromatin dynamics are distinct from active dynamics, and reveal hierarchical structure



Full model is complex, but integration simplifies the update equation for Gibbs sampling

$$\text{Prob}\left(\vec{z}, \vec{w}, \vec{\theta}, \vec{p} | \alpha, \lambda\right) = \text{Prob}\left(\text{topic}\right) \text{Prob}\left(\text{cell}\right) \text{Prob}\left(\text{genomic location} | \text{topic}\right)$$

Full model is complex, but integration simplifies the update equation for Gibbs sampling

$$\text{Prob}\left(\vec{z}, \vec{w}, \vec{\theta}, \vec{p} | \alpha, \lambda\right) = \text{Prob}\left(\text{topic}\right) \text{Prob}\left(\text{cell}\right) \text{Prob}\left(\text{genomic location} | \text{topic}\right)$$

$$\text{Prob}\left(\vec{z}, \vec{w}, \vec{\theta}, \vec{p} | \alpha, \lambda\right) = \prod_{k=1}^{K} \text{Prob}\left(p_k | \lambda\right) \prod_{d=1}^{D} \text{Prob}\left(\theta_d | \, \alpha\right) \prod_{n=1}^{N} \text{Prob}\left(z_{d,n} | \theta\right) \text{Prob}\left(w_{d,n} | \lambda_{z_{d,n}}\right)$$

Full model is complex, but integration simplifies the update equation for Gibbs sampling

$$\text{Prob}\left(\vec{z}, \vec{w}, \vec{\theta}, \vec{p} | \alpha, \lambda\right) = \text{Prob}\left(\text{topic}\right) \text{Prob}\left(\text{cell}\right) \text{Prob}\left(\text{genomic location} | \text{topic}\right)$$

$$\text{Prob}\left(\vec{z}, \vec{w}, \vec{\theta}, \vec{p} | \alpha, \lambda\right) = \prod_{k=1}^{K} \text{Prob}\left(p_k | \lambda\right) \prod_{d=1}^{D} \text{Prob}\left(\theta_d | \, \alpha\right) \prod_{n=1}^{N} \text{Prob}\left(z_{d,n} | \theta\right) \text{Prob}\left(w_{d,n} | \lambda_{z_{d,n}}\right)$$

$$\text{Prob}\left(\vec{z}, \vec{w} | \alpha, \lambda\right) = \int_{\vec{p}} \prod_{k=1}^{K} \text{Prob}\left(p_k | \lambda\right) \text{Prob}\left(z_{d,n} | \theta\right) \text{Prob}\left(w_{d,n} | \lambda_{z_{d,n}}\right) \int_{\vec{\theta}} \prod_{d=1}^{D} \text{Prob}\left(\theta_d | \, \alpha\right)$$

Collapsed Gibbs sampling updates efficiently

Probability of assigning read *n* to topic *k:*

Collapsed Gibbs sampling updates efficiently

Probability of assigning read *n* to topic *k:*

$$\text{Prob}\left(z_{d,n} = k | \vec{z}_{-d,n}, \vec{w}, \alpha, \lambda\right) = \frac{u_{d,k} + \alpha_k}{\sum_{k'}^{K} u_{d,k'} + \alpha_{k'}} \cdot \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_{w'}^{W} v_{k,w'} + \lambda_i}$$

$u_{d,k}$: number of times cell *d* uses topic *k*
$v_{k,w_{d,n}}$: number of times topic *k* uses locus $w_{d,n}$
α : Dirichlet prior for cell-to-topic distribution
λ : Dirichlet prior for topic-to-locus distribution

Collapsed Gibbs sampling updates efficiently

Probability of assigning read *n* to topic *k:*

$$\text{Prob}\left(z_{d,n} = k | \vec{z}_{-d,n}, \vec{w}, \alpha, \lambda\right) = \boxed{\frac{u_{d,k} + \alpha_k}{\sum_{k'}^{K} u_{d,k'} + \alpha_{k'}}} \cdot \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_{w'}^{W} v_{k,w'} + \lambda_i}$$

How much a cell likes a topic

$u_{d,k}$: number of times cell *d* uses topic *k*
$v_{k,w_{d,n}}$: number of times topic *k* uses locus $w_{d,n}$
α : Dirichlet prior for cell-to-topic distribution
λ : Dirichlet prior for topic-to-locus distribution

Collapsed Gibbs sampling updates efficiently

Probability of assigning read *n* to topic *k:*

$$\text{Prob}\left(z_{d,n} = k | \vec{z}_{-d,n}, \vec{w}, \alpha, \lambda\right) = \boxed{\frac{u_{d,k} + \alpha_k}{\sum_{k'}^{K} u_{d,k'} + \alpha_{k'}}} \cdot \boxed{\frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_{w'}^{W} v_{k,w'} + \lambda_i}}$$

How much a cell likes a topic

How much a topic likes a genomic locus

$u_{d,k}$: number of times cell *d* uses topic *k*
$v_{k,w_{d,n}}$: number of times topic *k* uses locus $w_{d,n}$
α : Dirichlet prior for cell-to-topic distribution
λ : Dirichlet prior for topic-to-locus distribution