UNIVERSITÄT DES SAARLANDES

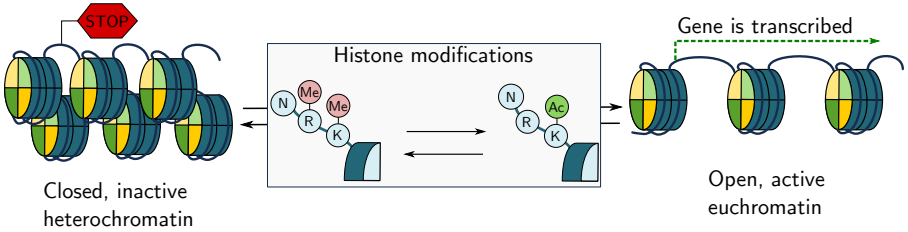
EpiSegMix

A Flexible Distribution Hidden Markov Model with Duration Modeling for Chromatin State Discovery

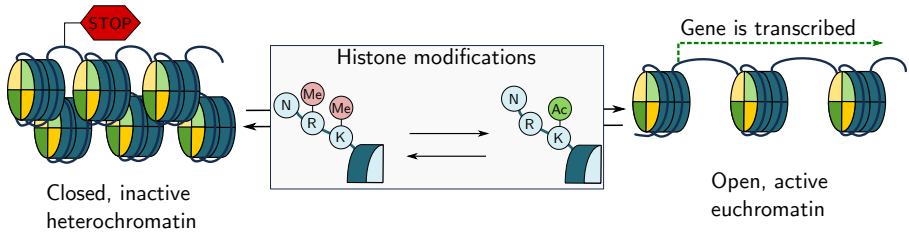
Johanna E. Schmitz, Nihit Aggarwal, Lukas Laufer, Jörn Walter, Abdulrahman Salhab, Sven Rahmann

DSB 2024

Biological Background

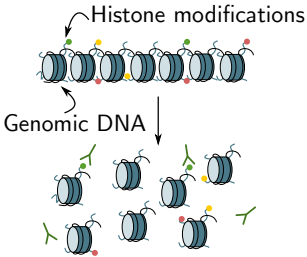


Biological Background

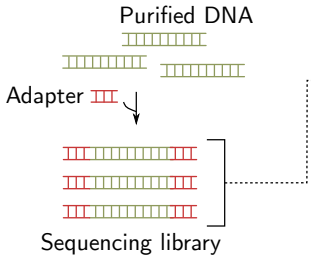


Modification	Function	Genomic Context
H3K9me3	Transcriptional repression	Heterochromatin
H3K27me3	Transcriptional repression	CpG-rich promoters, intergenic regions
H3K36me3	Transcriptional elongation	Gene bodies
H3K27ac	Transcriptional activation	Promoters, enhancers
H3K4me1	Transcriptional activation	Promoters, enhancers, intergenic regions
H3K4me3	Transcriptional activation	Promoters

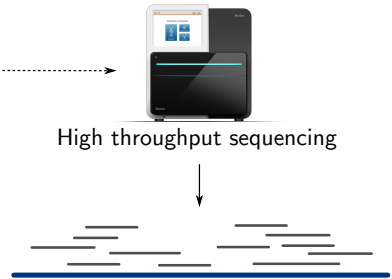
Chromatin Immunoprecipitation with DNA sequencing (ChIP-seq)



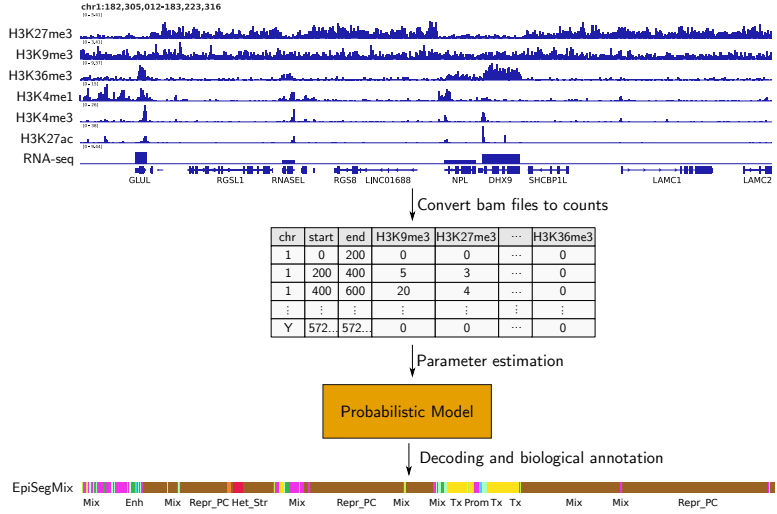
Crosslinking, fragmentation and immunoprecipitation



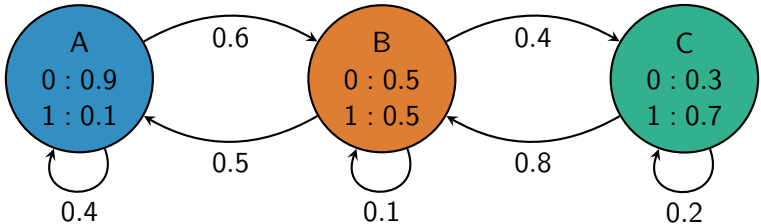
DNA purification and sequencing libary preparation



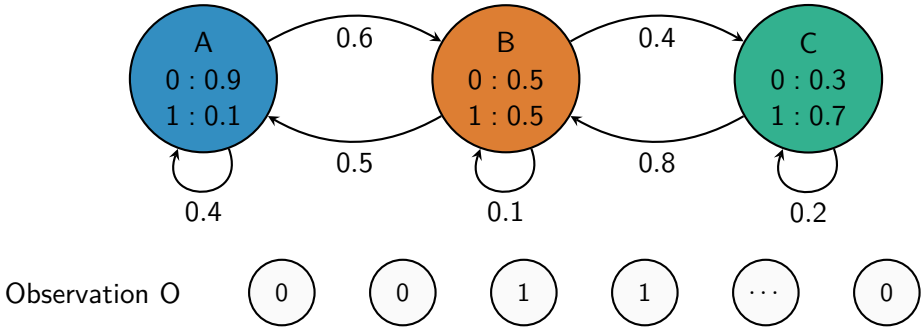
Chromatin Segmentation



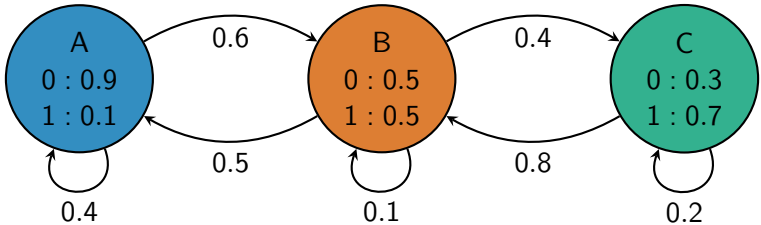
Introduction to Hidden Markov Models



Introduction to Hidden Markov Models



Introduction to Hidden Markov Models

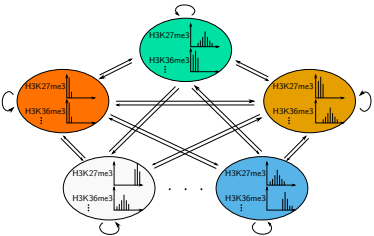


Hidden Markov Models for Chromatin Segmentation

1. Binned read counts

chr	start	end	H3K9me3	H3K27me3	···
1	0	200	0	0	···
1	200	400	5	3	···
1	400	600	20	4	···
1	600	800	9	0	···
1	800	1000	8	1	···
1	1000	1200	10	1	···
⋮	⋮	⋮	⋮	⋮	⋮

2. Probabilistic Model



3. Segmentation

chr	start	end	State
1	0	200	1
1	200	400	2
1	400	600	3
1	600	800	2
1	800	1000	2
1	1000	1200	2
⋮	⋮	⋮	⋮

ChromHMM[1]

- fully connected HMM
- read count distribution:
product of independent Bernoulli experiments
(binarized input using Poisson distribution)

EpiCSeg[2]

- fully connected HMM
- read count distribution:
multivariate negative binomial distribution

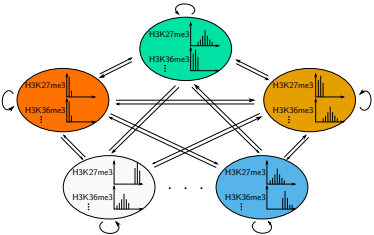
1 Ernst, Kellis (2010)
2 Mammana, Chung (2015)

Hidden Markov Models for Chromatin Segmentation

1. Binned read counts

chr	start	end	H3K9me3	H3K27me3	···
1	0	200	0	0	···
1	200	400	5	3	···
1	400	600	20	4	···
1	600	800	9	0	···
1	800	1000	8	1	···
1	1000	1200	10	1	···
⋮	⋮	⋮	⋮	⋮	⋮

2. Probabilistic Model



3. Segmentation

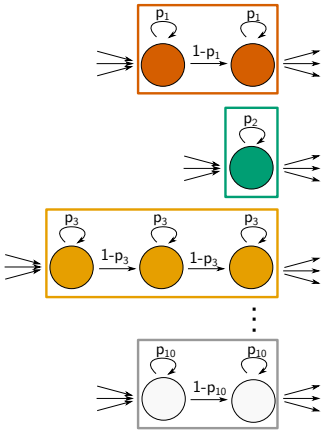
chr	start	end	State
1	0	200	1
1	200	400	2
1	400	600	3
1	600	800	2
1	800	1000	2
1	1000	1200	2
⋮	⋮	⋮	⋮

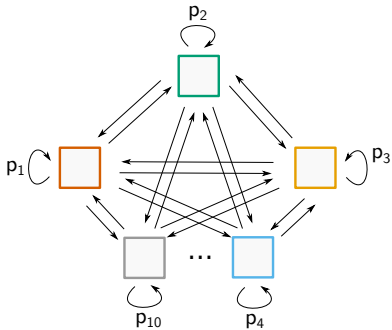
EpiSegMix

- extended state HMM
- read count distribution: configurable combination of discrete distribution types

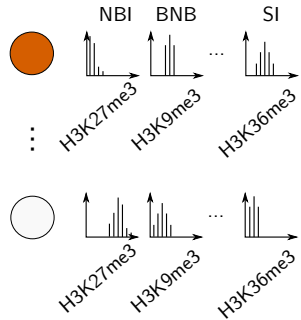
Flexible Modeling with EpiSegMix

Duration modeling





Emission modeling

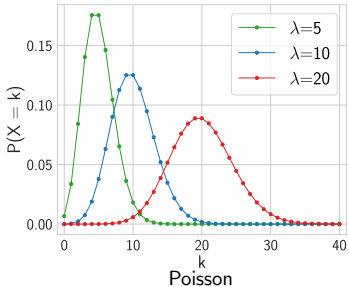


Flexible Emission Modeling

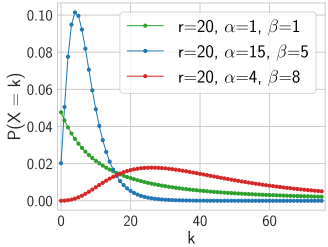
Table: Overview of available discrete distributions (Johnson et al., 1993).

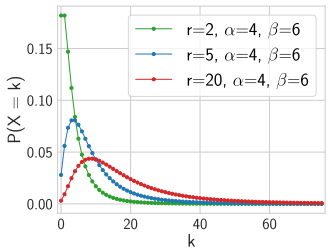
Name	Parameters	Flexibility
Poisson	$\lambda \in \mathbb{R}^+$	$+$
Binomial	$n \in \mathbb{N}, p \in [0, 1]$	$+$
Negative Binomial	$r \in \mathbb{R}^+, p \in [0, 1]$	$++$
Beta Binomial	$n \in \mathbb{N}, \alpha \in \mathbb{R}^+, \beta \in \mathbb{R}^+$	$+++$
Beta Negative Binomial	$r \in \mathbb{R}^+, \alpha \in \mathbb{R}^+, \beta \in \mathbb{R}^+$	$+++$
Sichel	$\mu \in \mathbb{R}^+, \sigma \in \mathbb{R}^+, \nu \in \mathbb{R}$	$+++$

Comparison Poisson and Beta Negative Binomial Distribution



Poisson

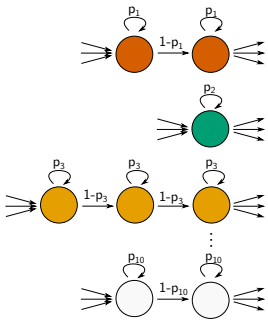




Beta Negative Binomial

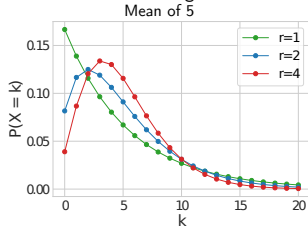
Flexible Duration Modeling

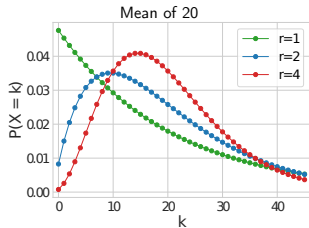
Left-right topology[1]



Sum of Geometric distributions
→ Negative Binomial distribution

Theoretical state length distribution





1 Russell, Cook (1987)

Parameter Estimation

HMM parameters

- Transition probabilities
- Emission probabilities
  $\rightarrow$ parameters of discrete distributions

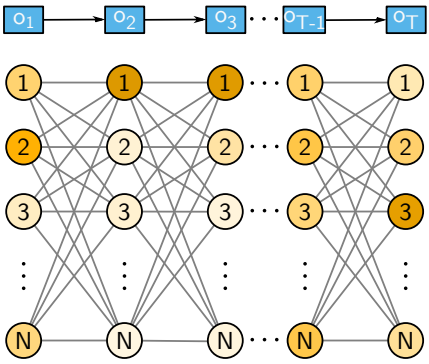
Parameter Estimation

HMM parameters

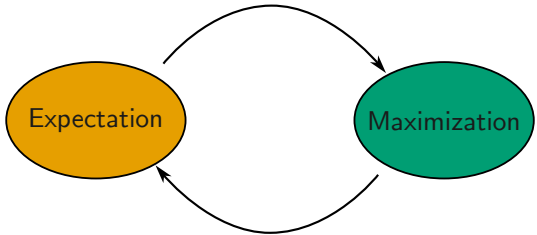
- Transition probabilities
- Emission probabilities
  $\rightarrow$ parameters of discrete distributions

Problem

- No mapping between observations and states available



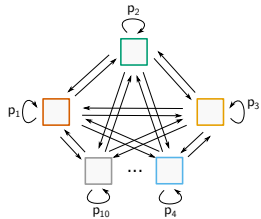
Parameter Estimation

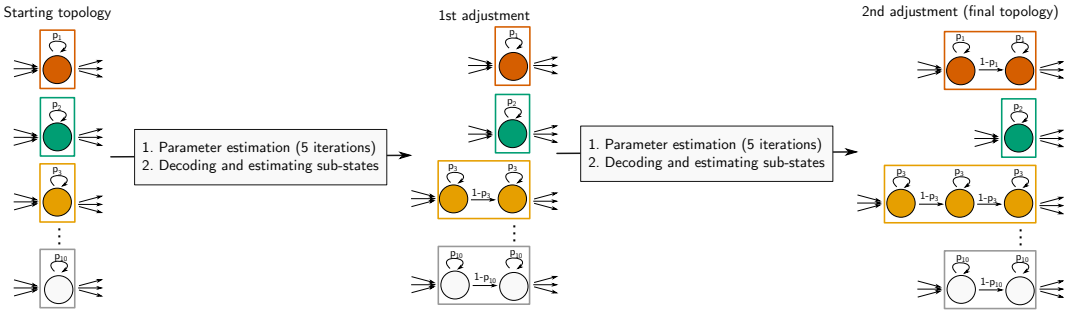


Solution

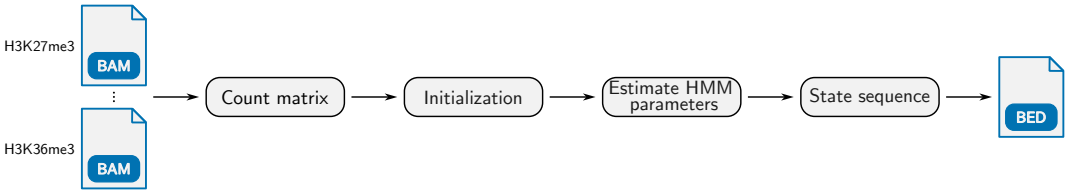
- Baum-Welch algorithm (expectation-maximization):
  - E-step: estimate state membership probabilities
  - M-step: estimate model parameters
- Emission parameters are updated by maximum likelihood

Parameter Estimation

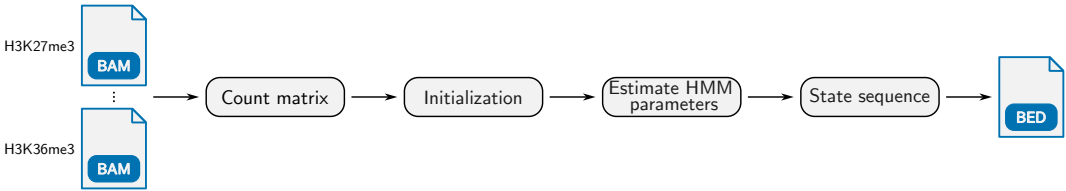




EpiSegMix Workflow

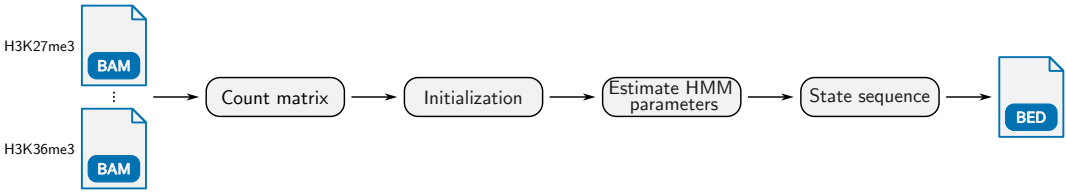


EpiSegMix Workflow

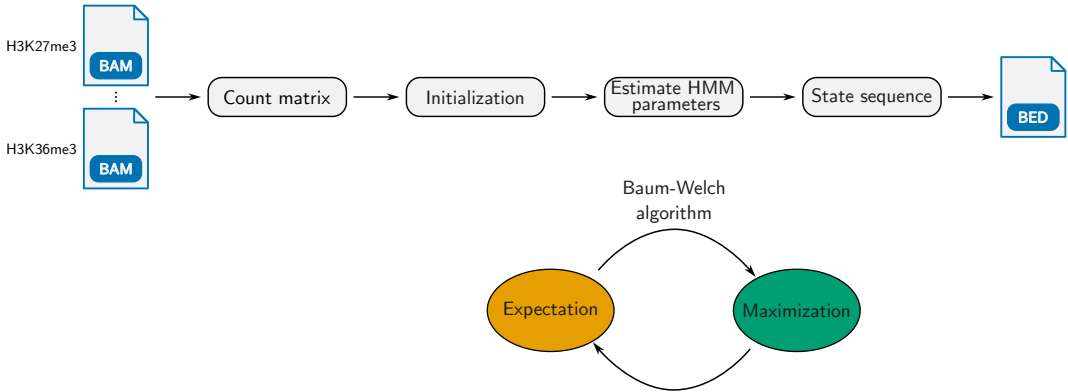


chr	start	end	H3K27me3	···	H3K36me3
1	0	200	0	···	0
1	200	400	5	···	10
1	400	600	20	···	1
⋮	⋮	⋮	⋮	⋮	⋮

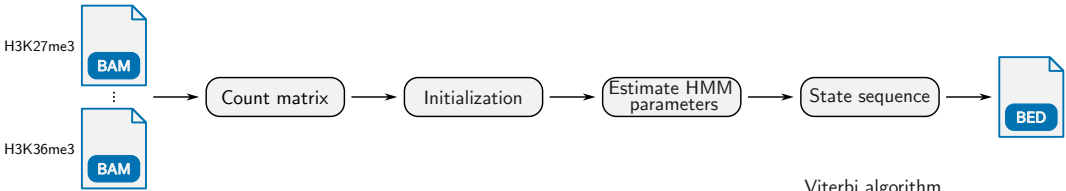
EpiSegMix Workflow



EpiSegMix Workflow

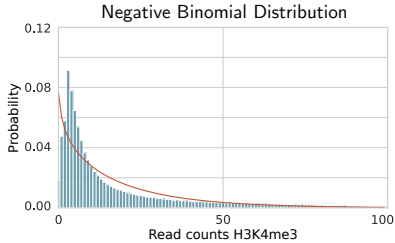


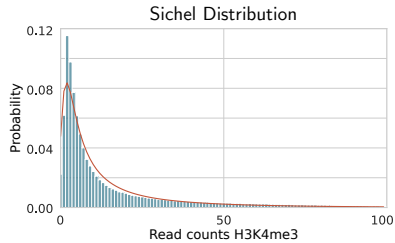
EpiSegMix Workflow



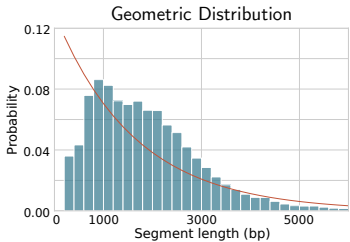
Results - Advantage of Flexible Modeling

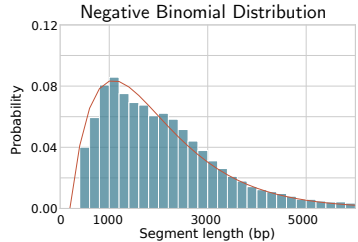
Emission Modeling



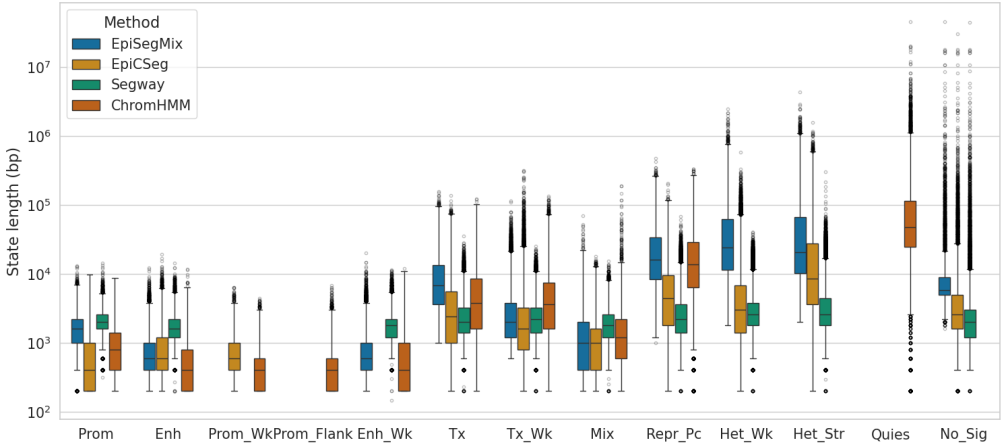


Duration Modeling





Results - Increased Segment Length



Results - EpiSegMix is Predictive of Cell Biology

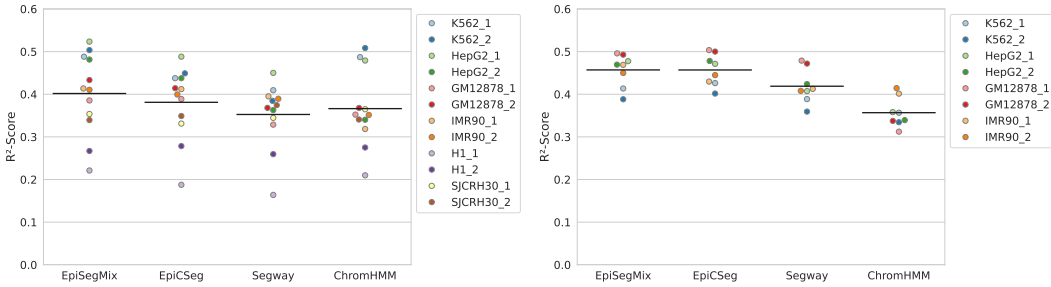
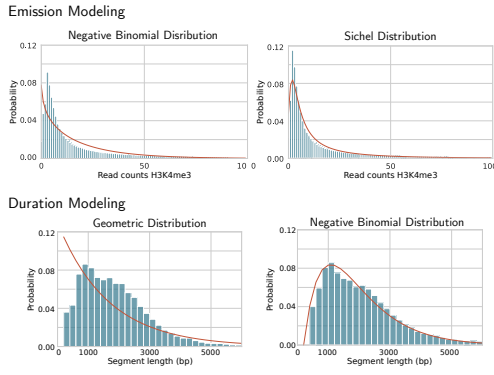


Figure: Prediction of (a) gene expression (b) ATAC-seq counts using state labels as categorical predictors.

Conclusion

- Automated chromatin state discovery enables the annotation of both coding and non-coding regions of the genome.
- The assumptions of the probabilitic model have an impact on the segmentation and biological interpretability of states.
- EpiSegMix provides a flexible framework that can be applied to read count data with different underlying distribution types.



https://gitlab.com/rahmannlab/episegmix

References I

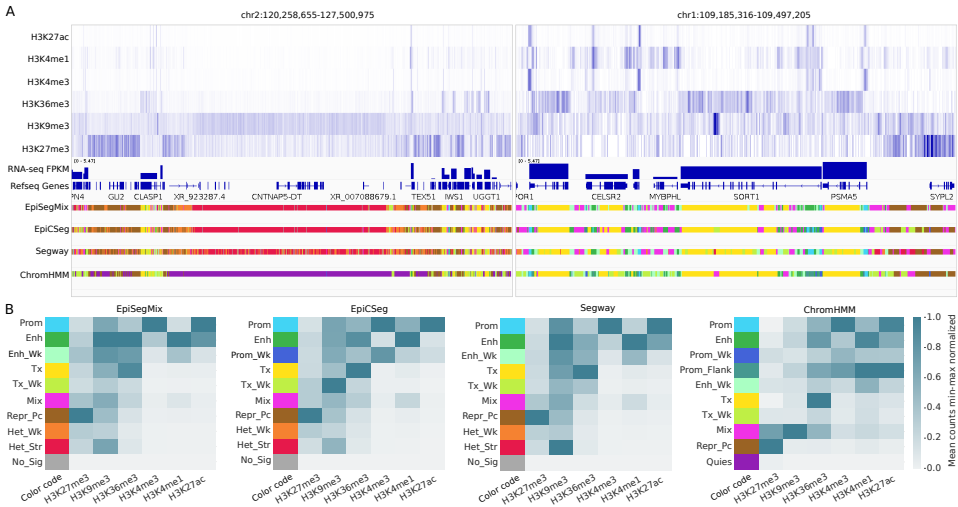
Ernst Jason, Kellis Manolis. Discovery and characterization of chromatin states for systematic annotation of the human genome // Nature Biotechnology. VIII 2010. 28, 8.

Johnson Norman L., Kotz Samuel, Kemp Adrienne W., others . Univariate discrete distributions. New York, NY [u.a.]: Wiley, 1993. 2nd ed.

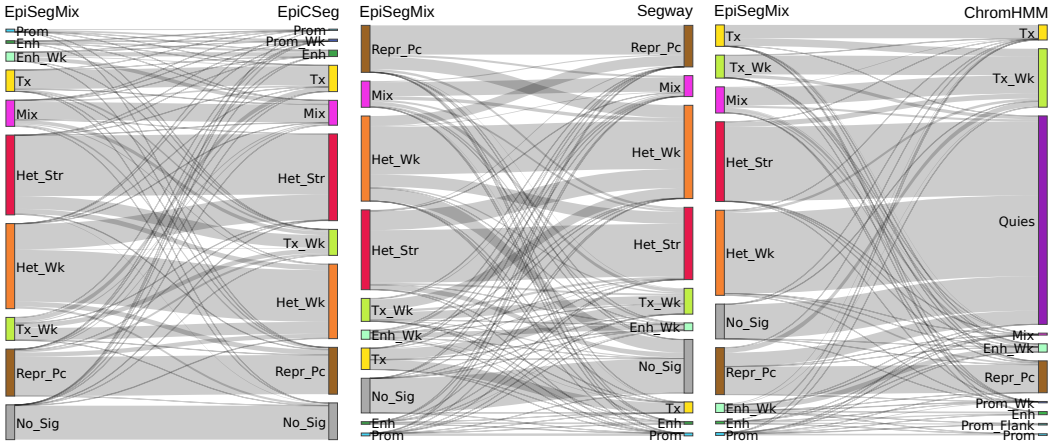
Mammana Alessandro, Chung Ho-Ryun. Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome // Genome Biology. VII 2015. 16, 1. 151.

Russell M., Cook A. Experimental evaluation of duration modelling techniques for automatic speech recognition // ICASSP '87. IEEE International Conference on Acoustics, Speech, and Signal Processing. 12. IV 1987. 2376–2379.

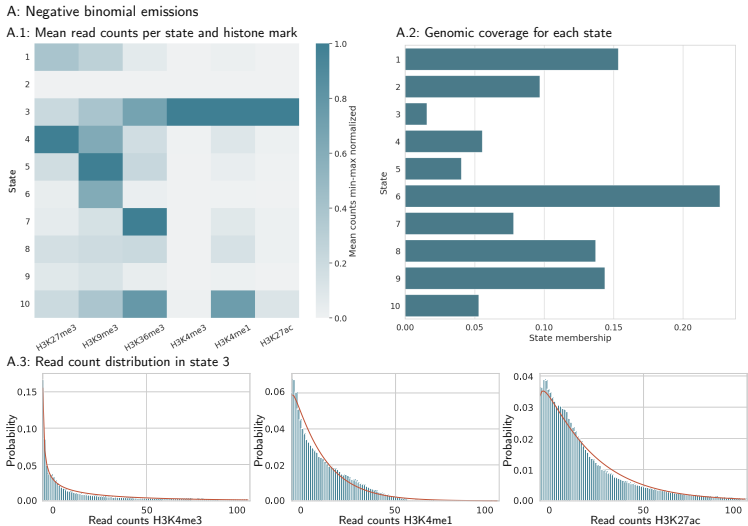
Appendix


A
chr2:120,258,655-127,500,975
chr1:109,185,316-109,497,205
H3K27ac
H3K4me1
H3K4me3
H3K36me3
H3K9me3
H3K27me3
RNA-seq FPKM
Refseq Genes
EpiSegMix
EpiCSeg
Segway
ChromHMM
B
EpiSegMix
EpiCSeg
Segway
ChromHMM
Prom
Enh
Enh_Wk
Tx
Tx_Wk
Mix
Repr_Pc
Het_Wk
Het_Str
No_Sig
Prom_Wk
Prom_Flank
Quies
Color code
H3K27me3
H3K9me3
H3K36me3
H3K4me3
H3K4me1
H3K27ac
Mean counts min-max normalized

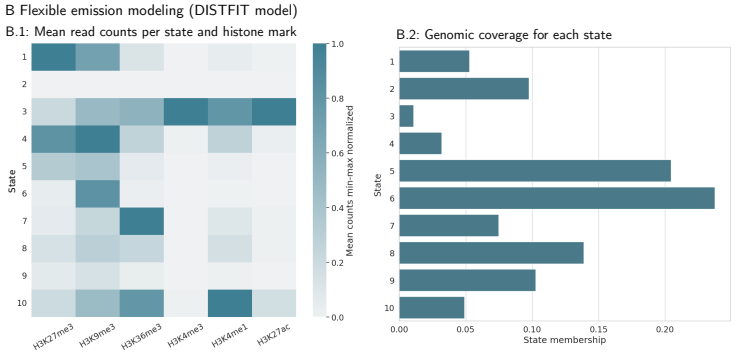
Appendix


EpiSegMix
EpiCSeg
Prom
Enh
Enh_Wk
Tx
Mix
Het_Str
Het_Wk
Tx_Wk
Repr_Pc
No_Sig
Prom
Prom_Wk
Enh
Tx
Mix
Het_Str
Tx_Wk
Het_Wk
Repr_Pc
No_Sig
EpiSegMix
Segway
Repr_Pc
Mix
Het_Wk
Het_Str
Tx_Wk
Enh_Wk
Tx
No_Sig
Enh
Prom
Repr_Pc
Mix
Het_Wk
Het_Str
Tx_Wk
Enh_Wk
No_Sig
Tx
Enh
Prom
EpiSegMix
ChromHMM
Tx
Tx_Wk
Mix
Het_Str
Het_Wk
No_Sig
Repr_Pc
Enh_Wk
Enh
Prom
Tx
Tx_Wk
Quies
Mix
Enh_Wk
Repr_Pc
Prom_Wk
Enh
Prom_Flank
Prom

Appendix


A: Negative binomial emissions
A.1: Mean read counts per state and histone mark
State
Mean counts min-max normalized
H3K27me3
H3K9me3
H3K36me3
H3K4me3
H3K4me1
H3K27ac
A.2: Genomic coverage for each state
State membership
A.3: Read count distribution in state 3
Probability
Read counts H3K4me3
Read counts H3K4me1
Read counts H3K27ac

Appendix


B Flexible emission modeling (DISTFIT model)
B.1: Mean read counts per state and histone mark
State
Mean counts min-max normalized
H3K27me3
H3K9me3
H3K36me3
H3K4me3
H3K4me1
H3K27ac
B.2: Genomic coverage for each state
State membership
B.3: Read count distribution in state 3
Probability
Read counts H3K4me3
Read counts H3K4me1
Read counts H3K27ac