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

Johanna Elena Schmitz [1,2,3†], Nihit Aggarwal [4†], Lukas Laufer [4], Jörn Walter [4], Abdulrahman Salhab[4,5‡], and Sven Rahmann[1,2‡]

[1]Algorithmic Bioinformatics, Center for Bioinformatics Saar, Saarland Informatics Campus, 66123, Saarbrücken, Germany
[2]Department MI, Saarland University, Saarland Informatics Campus, 66123, Saarbrücken, Germany
[3]Saarbrücken Graduate School of Computer Science, Saarland Informatics Campus, 66123, Saarbrücken, Germany
[4]Department of Genetics, Saarland University, 66123, Saarland, Germany
[5]Genomics Data Science Core, Integrated Genomics Services, Sidra Medicine, Doha, Qatar
[†]Contributed equally.
[‡]Contributed equally.

Abstract

Motivation: Automated chromatin segmentation based on ChIP-seq data reveals insights into the epigenetic regulation of chromatin accessibility. Existing segmentation methods are constrained by simplifying modeling assumptions, which may have a negative impact on the segmentation quality. We introduce EpiSegMix, a segmentation tool based on a hidden Markov model with flexible read count distribution types and state duration modeling, allowing for a more flexible modeling of both histone signals and segment lengths.

Results: In a comparison with two existing tools, ChromHMM and EpiCSeg, we show that EpiSegMix is more robust to technical variability. Furthermore, its flexible framework enables it to fit an accurate probabilistic model, which has the potential to increase the biological interpretability of chromatin states.

Availability and implementation: Source code is available at `https://gitlab.com/rahmannlab/episegmix`.

Corresponding author: Sven Rahmann

1 Introduction

Each cell in a eukaryotic organism contains the same genetic information to build all required structural and functional gene products. However, cell-to-cell variation is essential for having specialized tissues with distinct physiological functions and to adapt to environmental changes (Cavalli and Heard, 2019; Carter and Zhao, 2021). This necessitates an additional layer of processes regulating gene expression to enable cell differentiation and to maintain cellular identities throughout cell divisions (Allis and Jenuwein, 2016). Among the mechanisms tightly regulating gene expression are transcription factors and epigenetic modifications, like DNA methylation and histone modifications.

Histone modifications With the increasing knowledge about the role of histone modifications in altering the chromatin structure, and hence DNA accessibility, it became apparent that different histone modifications are predominantly enriched in chromatin regions with distinct functional roles (Baker, 2011). For example, H3K4me3 (trimethylation (**me3**) of lysine at position 4 (**K4**) in the amino acid sequence of the histone protein **H3**) is predominantly enriched in promoters and can be linked to transcriptional activation, H3K36me3 is enriched in active genes, and H3K27me3 can be associated with gene repression by the Polycomb protein complex (Blackledge and Klose, 2021). Furthermore, densely packed chromatin, called heterochromatin, is typically characterized by low levels of acetylation, whereas open, actively transcribed chromatin, called euchromatin, shows an enrichment of acetylated lysine (Bannister and Kouzarides, 2011). Therefore, the combinatorial patterns of multiple histone modifications allow us to characterize so-called chromatin states that describe the different functional states of both coding and non-coding regions in the genome (Baker, 2011).

ChIP-seq Advances in chromatin immunoprecipitation followed by sequencing (ChIP-seq) boosted the understanding of the various roles of histone modifications by providing the ability to generate genome-wide histone maps in high-throughput experiments (Barski et al., 2007). For ChIP-seq, DNA-binding proteins of interest, such as specifically modified histones or transcription factors, are tagged with specific antibodies. After shearing the DNA, the DNA fragments that are bound to the desired proteins can be isolated and extracted. This process is not completely specific, and a certain amount of noise or background DNA has to be expected. The DNA molecules are then purified and sequenced, and after mapping the se-

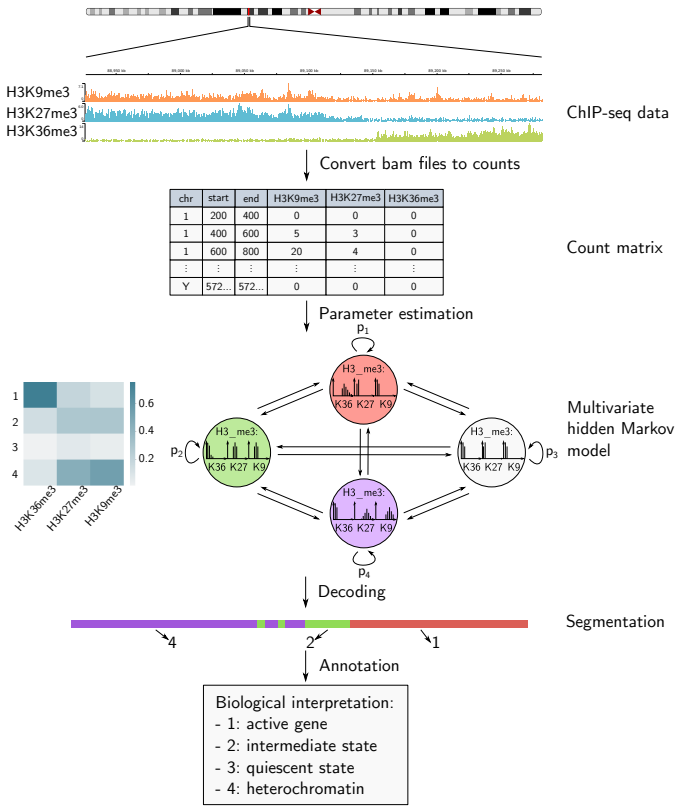


Figure 1: Chromatin segmentation. The reads of a ChIP-seq experiment are converted into a count matrix by counting the number of reads mapping to each non-overlapping 200bp genomic interval. The states of a multivariate Hidden Markov model (HMM) capture patterns in the multivariate read count distribution of the histone marks, and the transition probabilities between states capture the relations between adjacent chromatin states.

quenced reads to the reference genome, we can infer which histone marks occur in which parts of the genome (Park, 2009) using a probabilistic model.

Probabilistic models for segmentation The availability of genome-wide ChIP-seq data led to the development of automated methods for genome segmentation and annotation. These methods all share the common idea of exploiting a probabilistic model to detect recurrent patterns of epigenetic marks using the aligned reads to determine the signal intensity at different positions in the genome (see Figure 1). For example, HMMSeg (Day et al., 2007), ChromHMM (Ernst and Kellis, 2010) and EpiCSeg (Mammana and Chung, 2015) are based on hidden Markov models (HMM), Segway and Segway 2.0 (Chan et al., 2018) fit a Gaussian mixture model using a dynamic Bayesian network, and Daneshpajouh et al. (2022) developed a state–space model assuming that the observed data is a linear function in the state-specific parameter matrix plus Gaussian noise. For a more comprehensive review of existing methods, we refer the reader to the review by Libbrecht et al. (2021).

HMMs Multivariate HMMs are able to capture both combinatorial patterns of multiple histone marks and spatial relations between different genomic elements, which makes them a prominent tool for chromatin state discovery (Lee and Park, 2014). An HMM describes two stochastic processes, an invisible Markov chain consisting of a finite set of hidden states and a visible process of observable signals. Here, the hidden states correspond to chromatin states and the observable signals to the observed read counts per genomic region. Each hidden state has state-specific probabilities of emitting an observation, called *emission probabilities*. For chromatin segmentation, the genome is divided into non-overlapping intervals (of typical length 200 bp), such that each observation is a vector of counts corresponding to the number of reads assigned to the interval per histone mark. Thus, the emission probabilities of a single hidden state capture a specific combinatorial pattern of multiple histone marks. Different states may hence define different functional genomic elements, such as promoters, enhancers or gene bodies. In addition, the relations between adjacent chromatin states are modeled via transition probabilities, which determine, for each state, the probabilities to either stay in the same state or to transition to another state.

Modeling assumptions Since the prospect of an HMM to discover biologically meaningful chromatin states is to a large extent depending on its ability to find recurrent patterns in the data, the modeling assumptions underlying the emission probabilities are of particular importance and are thus the distinguishing feature of existing methods. For example, ChromHMM fits an HMM on binarized data (high vs. low read count), where the emission probabilities are assumed to be independent Bernoulli experiments, and EpiCSeg models the emission probabilities using a negative multinomial distribution. However, previous analyses of ChIP-seq data have shown that the read count distributions are often overdispersed and skewed, partly caused by differential protection against sonication, unequal binding affinity of distinct antibodies, sequence dependent PCR amplification and discrepancies when mapping to repeat-rich regions, which all introduce bias to the data (Diaz et al., 2012). Comparing the read count histograms of different histone modifications on different datasets reveals that some histone modification are more prone to overdispersion and skewness, which cannot be accounted for by some of the commonly used distributions, such as the negative binomial distribution. Furthermore, Beacon et al. (2021) showed that histone marks that are enriched in short domains, like promoters or TSS, are typically characterized by very narrow peaks with high signal intensities, while histone marks enriched in broad domains, like heterochromatic regions, have lower signal intensities. This has the additional consequence that contiguous regions in the genome covered by the same chromatin state have different lengths, for example, short promoters and long heterochromatic regions. However, existing tools are limited to a single duration distribution type (geometric, i.e. exponentially decreasing probabilities) and only vary the mean length.

Name	Parameters	Reference
Binomial	$n \in \mathbb{N}, p \in [0,1]$	Bernoulli (1713)
Poisson	$\lambda \in \mathbb{R}^+$	Poisson (1837)
Negative binomial	$r \in \mathbb{R}^+, p \in [0,1]$	Lindén and Mäntyniemi (2011)
Beta binomial	$n \in \mathbb{N}, \alpha \in \mathbb{R}^+, \beta \in \mathbb{R}^+$	Griffiths (1973)
Beta negative binomial	$r \in \mathbb{R}^+, \alpha \in \mathbb{R}^+, \beta \in \mathbb{R}^+$	Wang (2011)
Sichel	$\mu \in \mathbb{R}^+, \sigma \in \mathbb{R}^+, v \in \mathbb{R}$	Sichel (1992)

Table 1: Overview of available discrete distributions.

Novel contributions We propose a new flexible and highly configurable HMM that generalizes the modeling assumptions of existing HMM-based methods in two ways. First, we are able to model overdispersed skewed read count distributions by providing the option to choose, for each histone modification, a different discrete distribution from a broad selection (Table 1). Second, we propose to use an extended-state HMM with flexible duration modeling to better capture the characteristics of broad and narrow chromatin domains.

By applying our method to publicly available ChIP-seq data, we show that such a flexible HMM reduces technical variability in the segmentation and has the potential to increase the segmentation accuracy and biological interpretability, compared to existing methods.

2 Methods

2.1 Probabilistic Model

A hidden Markov model (HMM) is formally defined as a quintuple (N, π, Σ, A, B), where $\{1, 2, \ldots, N\}$ is a finite set of hidden states, π is a probability vector with the starting probabilities for each state, Σ is a set of possible output values, A is an $N \times N$ matrix where each entry a_{ij} denotes the transition probability to move from state i to state j, and $B = (b_j(\cdot) \mid j \in \{1, 2, \ldots, N\})$ is the complete collection of parameters required to calculate the emission probabilities of an observation in each state j (Lee and Park, 2014). For a sample with T observations or time points, the collection of random variables is thus given by (O, Q), where $O = (O_1, \ldots, O_T)$ denotes the observed sequence and $Q = (Q_1, \ldots, Q_T)$ denotes the hidden state sequence.

Due to the HMM independence assumptions (Q is a Markov chain, O_t is conditionally independent of everything else, given Q_t; Bilmes (1998)), the probability that an HMM with parameters $\theta = (\pi, A, B)$ generates the combination $(Q = q, O = o)$ is

$$P_\theta(Q = q, O = o) = \pi_{q_0} \prod_{t=1}^{T} \alpha_{q_{t-1} q_t} b_{q_t}(o_t) \,. \tag{1}$$

For given fixed $O = o$, the Viterbi algorithm (Rabiner, 1989) allows us to determine the state sequence q that maximizes this probability (for fixed transition and emission probabilities). Alternatively, posterior decoding finds the state sequence which maximizes the marginal posterior probability $P_\theta(Q_t = q_t \mid O = o)$ for each time point t separately.

For chromatin segmentation, the emission alphabet is given by the multivariate, countable infinite set

$$\Sigma = \{(x_1, \ldots, x_M) \mid x_m \in \mathbb{N}_0,\, m \in \{1, \ldots, M\}\} \,, \tag{2}$$

where M is the dimension of the emission alphabet, corresponding to the number of histone modifications in the input data. Hence, the natural choice is to model the emission probabilities using a multivariate discrete distribution. Under the assumption that the read counts of all histone modifications are conditionally independent of each other given a state, the emission probability for an observation $o_t = (o_{t1}, \ldots, o_{tM})$ in state j is given by

$$\begin{aligned} b_j(o_t) &= P_\theta(O_t = o_t \mid Q_t = j) && (3)\\ &= P_\theta(O_{t1} = o_{t1}, \ldots, O_{tM} = o_{tM} \mid Q_t = j) && (4)\\ &= \prod_{m=1}^{M} P_\theta(O_{tm} = o_{tm} \mid Q_t = j) \,. && (5) \end{aligned}$$

This leads to a flexible framework in which different distribution types may be selected to model the read counts of distinct histone marks.

Table 1 gives an overview over the univariate discrete distributions available in EpiSegMix (see Supplementary Methods for details). Distributions with more parameters are more flexible and hence lead to a more accurate model fit: One-parameter distributions (Poisson) may fit the mean of the observed read counts, but not variance or skewness; two-parameter distributions (e.g., negative binomial) may fit both mean and variance but not skewness; three-parameter distributions may fit all three moments. With data from several thousand genomic intervals per state, there is no danger of overfitting three parameters. Still, each distribution has its own limitations and dependencies between moments, so having a variety of options is beneficial.

2.2 Duration Modeling

The classical HMM topology is a fully connected graph, hence the state duration follows a geometric distribution with failure probability given by the self-transition probability of a state (Rabiner, 1989). Although the mean of a geometric distribution can be made arbitrarily large,

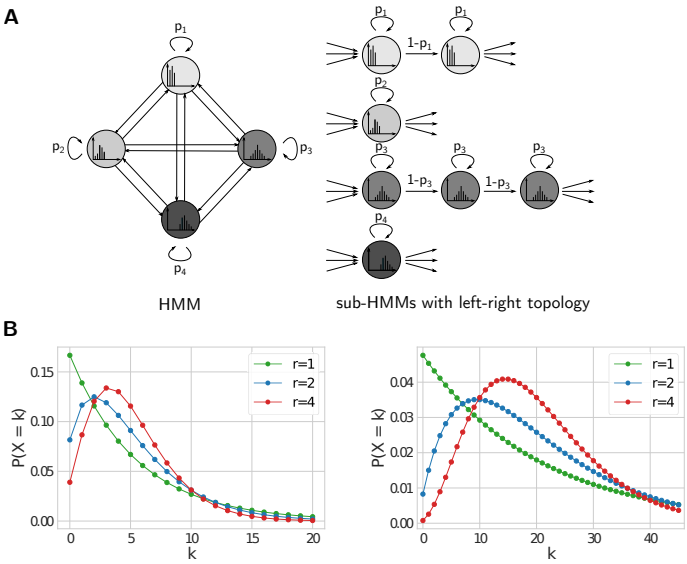


Figure 2: Duration modeling. **A** Extended-state HMM with different numbers of states in the sub-HMMs for a univariate 4-state HMM. **B** Comparison of state duration distribution for a state in a classical HMM ($r = 1$) and in a topology HMM with 2 or 4 sub-states to achieve the same mean of 5 (left) or 20 (right).

the mode is always at $P(X = 0)$, while the variance increases. Thus, geometric distributions are particularly well-suited to model shorter state durations, but have a limited flexibility when modeling longer durations since a higher mean results in a higher variance. We therefore propose to use an extended-state HMM, in which each state is internally represented as a sub-HMM. Each sub-HMM has a linear left-to-right topology with a different number of sub-states, but with the additional constraints that all sub-states have the same emission probabilities and same self-transition probabilities (Russell and Cook, 1987), see Figure 2A. Under this topology, the state duration follows a negative binomial distribution with parameters equal to the self-transition probability and to the number of sub-states. In contrast to a geometric distribution, the variance of a negative binomial distribution decreases for the same mean with an increasing number of sub-states (see Figure 2B). Therefore, both long and short durations can be modeled more flexibly.

To determine a suitable number of states for each sub-HMM, the segment lengths are computed for each state under the current model. The number of sub-states are then given by the estimated parameter r of a negative binomial distribution under the additional constraint that r is a natural number between one and five. By default, the topology is adjusted twice. Between the adjustments, the parameters of the HMM are estimated for a fixed number of iterations (default 5) and for the final topology until convergence.

2.3 Parameter Estimation

Before an HMM can be used to estimate the state sequence, its parameters (transition and emission probabilities) must be estimated. As we typically do not have labeled training data, parameter estimation must proceed in an unsupervised manner using the Baum-Welch algorithm, which is a concretization of the general expectation maximization (EM) algorithm tailored to the structure of HMMs (Rabiner, 1989). It alternates between estimating the model parameters (M-step), given a (fuzzy or probabilistic) assignment of observations to states, and re-estimating the state membership probabilities of each observation (E-step) until convergence. To guarantee the convergence of the Baum-Welch algorithm to a local optimum, the maximum likelihood estimators (MLE) are used to update the parameters of all distributions. If the MLE of a distribution has no closed form solution, it is computed numerically by optimizing the log-likelihood function (see Supplementary Methods for details). As an iterative technique, the Baum-Welch algorithm is sensitive to the given initial parameters. Therefore, systematically guessed initial values are beneficial both for convergence speed and goodness-of-fit. To find proper starting values, we first perform k-means clustering with k equal to the desired number of states. In the next step, the method-of-moments estimators are calculated for each cluster and used to initialize the emission parameters. The transition probabilities are initialized uniformly.

3 Implementation

The flexible distribution HMM and parameter estimation is implemented as a command-line tool "EpiSegMix" in C++. All steps of the surrounding workflow are incorporated into a Snakemake workflow (Mölder et al., 2021).

By default, we estimate parameters on the Encode pilot regions which contain a good representation of the whole genome and are thus commonly used to fit the model (Hoffman et al., 2012; Daneshpajouh et al., 2022). Alternatively, the user may specify a list of chromosomes to be used for model fitting. The segmentation is performed genome-wide.

All computations are performed in log-space for numerical stability (Mann, 2006). Numerical optimization is performed using the *MIGRAD()* function from the *ROOT Minuit 2* package (Brun and Rademakers, 1997).

The main output of chromatin segmentation is a file that assigns each position in the genome one state, and an HTML report with plots that characterize the model and segmentation, hence enabling their biological interpretation. For example, the heatmap showing the normalized histone modification intensities of each state (such as in Figure 3B) is central to determine the genomic function of the states.

Source code is available at `https://gitlab.com/rahmannlab/episegmix`.

4 Results

We evaluated EpiSegMix on publicly available ChIP-seq data for K562, IMR90 and H1 cell lines provided by the

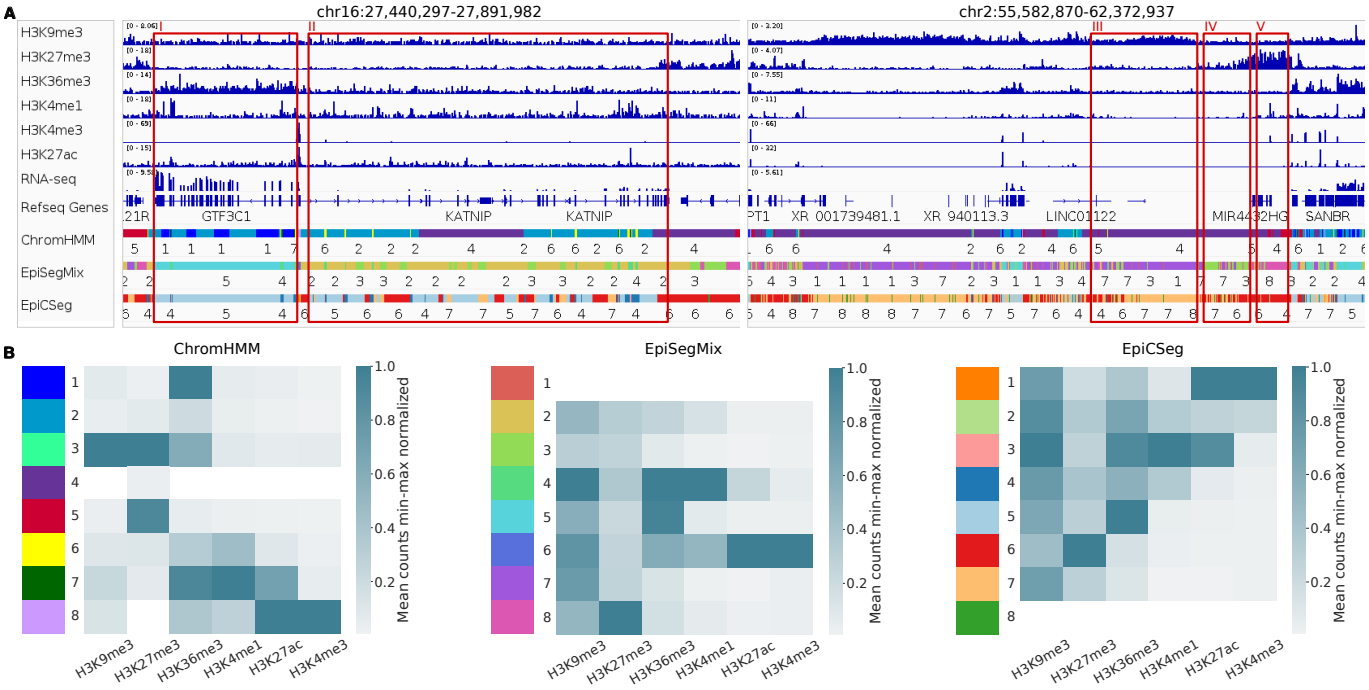


Figure 3: Segmentation comparison between tools for K562_1 using an 8-state HMM. **A** IGV-screenshot for two regions with different genomic activities. The first six tracks show the ChIP-seq signals of the six core histone marks. The 7th track shows the signal of unique reads from an RNA-seq experiment available from ENCODE (accession *ENCFF165DQJ*), and the 8-th track shows the Refseq gene positions. The last three tracks show the segmentation of ChromHMM, EpiSegMix and EpiCSeg. **B** Heatmaps showing the histone enrichment per state. The mean counts per state are normalized column wise, such that for each histone modification the state with the maximum mean count has a value of one and the state with the lowest mean count has a value of zero.

ENCODE consortium (Dunham et al., 2012). To analyze the robustness of all methods, we generated the count matrix for two replicate experiments of K562 and IMR90. To reproduce the results, all accession numbers are provided in Supplementary Table 2 and the script to create the count matrix is part of the code repository. We restricted our analysis to the six core histone marks H3K9me3, H3K27me3, H3K36me3, H3K4me1, H3K4me3 and H3K27ac defined by the IHEC consortium (Bujold et al., 2016).

Since chromatin segmentation is an unsupervised method and no ground truth is available, we evaluate the performance of EpiSegMix by intrinsic quality measures (robustness) and by comparing it to two established chromatin segmentation tools ChromHMM (Ernst and Kellis, 2012) and EpiCSeg (Mammana and Chung, 2015).

4.1 Data Processing

In a preprocessing step, we convert the aligned reads into a count matrix using the *bamsignals* package (Mammana and Helmuth, 2023). In the count matrix, each row corresponds to a consecutive, non-overlapping region with a fixed window size (default 200 base pairs), called bins, and each column corresponds to a distinct histone mark. Each read is then assigned to exactly one bin depending on the position of its 5' end.

For all models, we fitted an 8-state HMM. EpiSegMix was trained on the Encode pilot regions of *hg38* if not specified otherwise. For ChromHMM and EpiCSeg, the default parameters were used. For more information on the selected number of states and distributions for EpiSegMix, see Supplementary Figure 1 and Supplementary Table 1.

4.2 Exemplary Segmentation Comparison

We first show the impact of different modeling assumptions by highlighting the observed differences in two representative regions of the genome. One region is typical for an actively transcribed region and the other region shows a large heterochromatic domain (see Figure 3A, left and right respectively). In general, all methods find similar patterns of input histone marks including enrichment of regulatory marks such as H3K27ac, H3K4me1 and H3K4me3, actively transcribed mark H3K36me3, and repressive marks like H3K27me3 and H3K9me3. A closer look at the segmentation shows the advantages and disadvantages of the methods. The RNA-seq pattern of the two genes, GTF3C1 and KATNIP, shows that both genes have different expression levels in K562. Our EpiSegMix captures this difference accurately when compared to ChromHMM and EpiCSeg: It assigns different states, state 5 (highly expressed) for GTF3C1 and state 2 (lowly expressed) for KATNIP based mainly on the activity of H3K36me3 known for its association with transcription elongation along with other marks (see Figure 3B, I and II). By contrast, both EpiCSeg and ChromHMM fail to distinguish these regions and assign the same state to the

larger parts of the GTF3C1 and KATNIP genes.

When examining a large heterochromatic region, we see that EpiSegMix is able to model the transition from strong constitutive heterochromatic regions with high H3K9me3 (state 7) to Polycomb-associated heterochromatin with high H3K27me3 (state 8), see Figure 3A III, IV and V. This smooth transition is less visible for ChromHMM, which annotates almost the complete region as state 4. This state provides little or no insight as it shows a low enrichment for all marks (Figure 3B). EpiCSeg is capable of capturing different regions as EpiSegMix, but fails to provide precise boundaries between these transitions leading to a blurry segmentation track.

Overall, EpiSegMix provides well defined chromatin states that capture the functional genome biology.

4.3 Genome-wide Comparison

Figure 4A shows the genome-wide commonalities and differences of state assignments to genomic intervals among the three tools ChromHMM, EpiSegMix (ours) and EpiCSeg. The model states were manually relabeled with their best associated biological functions based on the enrichment of all histone marks per state (Figure 3B), the gene expression (Figure 4B) and the distribution of each state across 3000 bp up and down stream of the gene body (see Figure 4C). For example, in Figure 4C the promoter state in EpiSegMix (state 6) is characterized by high levels of both H3K27ac and H3K4me3 and is enriched at the gene start. Strong heterochromatic states show low occupancy around transcribed regions.

The majority of the genome is annotated with states that have a similar biological interpretation across different models, but the in-depth biological functions show differences. EpiSegMix accurately distinguishes between high and low transcribed regions across the genome in comparison to ChromHMM. EpiSegMix's transcription states, Tx_Str(5) and Tx_Wk(2) consist of 5745 and 3017 high and comparatively low expressed genes, respectively (see Figure 4B, middle). The two transcription-associated states (Tx_Str and Tx) defined by ChromHMM can partially distinguish these regions, but their gene expression overlaps, as depicted in Figure 4B. In contrast, EpiCSeg assigns all genes to its Tx state missing a distinction of high and low transcribed genes (see Figure 4A).

ChromHMM explains over 50% of the genome by a single state that is characterized by an overall low signal intensity and therefore provides no further biological insights. In contrast, EpiSegMix divides this state into a no signal state, a strong heterochromatic state and an intermediate state. The no signal and strong heterochromatic states cover gene-poor regions or genes that are lowly expressed. This is well observed in Figure 4C where its enrichment around and across protein coding genes is in between active and inactive states.

ChromHMM is expected to be more robust to noise due to the use of binarized values. However, the threshold for binarization is critical to call if a histone mark is present which may lead to oversimplification of the model, resulting in a large part of the genome being captured as a no-signal state. On the other hand, modeling the read counts using a more flexible multivariate distribution is more prone to fitting noise, which can lead to a fragmented segmentation as seen in EpiCSeg. In our model, we use duration modeling to overcome this limitation. Figure 4D shows that EpiSegMix is able to maintain longer segments compared to EpiCSeg. For instance, EpiSegMix's Het_Str(7) state has on average a higher segment length compared to the EpiCSeg's Het(7) state; hence EpiSegMix is able to better capture broad heterochromatic regions.

Similar results apply to the other datasets, which are included in the Supplementary Figures.

4.4 Robustness to Technical Variability

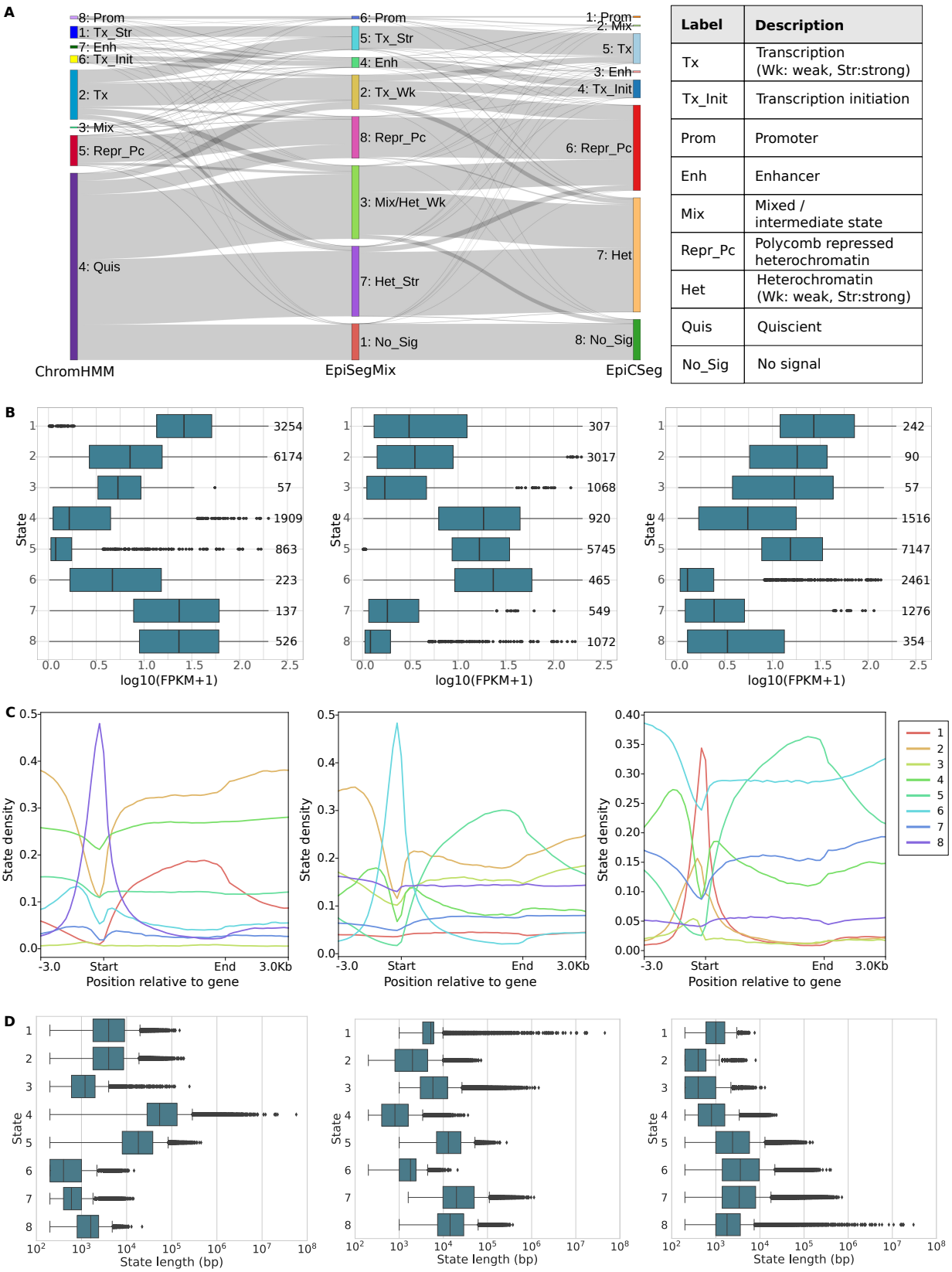
We evaluated the robustness of the three methods with regard to technical and biological variability using the isogenic replicates provided by ENCODE. To measure robustness, we computed the Jaccard index, averaged over all states, between the segmentations of the two replicates K562_1, K562_2 and IM90_1, IMR90_2, respectively, in the same way as done by Mammana and Chung (2015), i.e., model fitting and segmentation was performed separately on each chromosome and for each replicate; and states corresponding to each other were matched. The Jaccard index between two replicates is defined as the number of segments assigned the same state, divided by the total number of segments, and hence is between 0 (no agreement at all) and 1 (complete agreement). A higher index reflects higher robustness. On both cell lines K562 and IMR90, EpiSegMix has the highest average Jaccard index for most chromosomes, indicating that EpiSegMix is more robust to technical variability compared to ChromHMM and EpiCSeg (see Figure 5).

4.5 Computational Comparison

All three tools have comparable main memory requirements (3–5 GB of RAM) and computation times in the order of several minutes for the complete workflow. Computation time varies and mainly depends on the number of iterations until convergence of parameters, which is influenced by the (randomized) initialization. EpiSegMix (3.5 to 6 minutes) tends to be slightly slower than the other tools (ChromHMM: 4 to 5.5 minutes, EpiCSeq: 1.5 to 3.5 minutes), which may partially result from the additional flexibility and the (small) time spent on optimizing the code.

5 Discussion

We developed EpiSegMix, a flexible HMM for chromatin segmentation. We have increased the flexibility of the HMM with respect to modeling both emission probabilities and state durations. To account for the overdispersed and skewed ChIP-seq data, the read counts of each histone modification can follow a different discrete distribution. We currently support a variety of distributions, and further distributions can be added in the future. Furthermore, the internal HMM topology is adjusted to be able



Label	Description
Tx	Transcription (Wk: weak, Str:strong)
Tx_Init	Transcription initiation
Prom	Promoter
Enh	Enhancer
Mix	Mixed / intermediate state
Repr_Pc	Polycomb repressed heterochromatin
Het	Heterochromatin (Wk: weak, Str:strong)
Quis	Quiscient
No_Sig	No signal

Figure 4: State comparison between ChromHMM (left), EpiSegMix (middle) and EpiCSeg (right) for K562_1. **A** State overlap between methods. The bar widths corresponds to the genome coverage of each state and the edge widths to the overlap between states of different segmentations. The states are annotated with their biological function. **B** Gene expression. The plot shows the number of expressed genes and the expression levels covered by each state. **C** State distribution. The line plots show how often each state occurs around protein coding genes. **D** State duration. Each row shows for the respective state the segment length distribution.

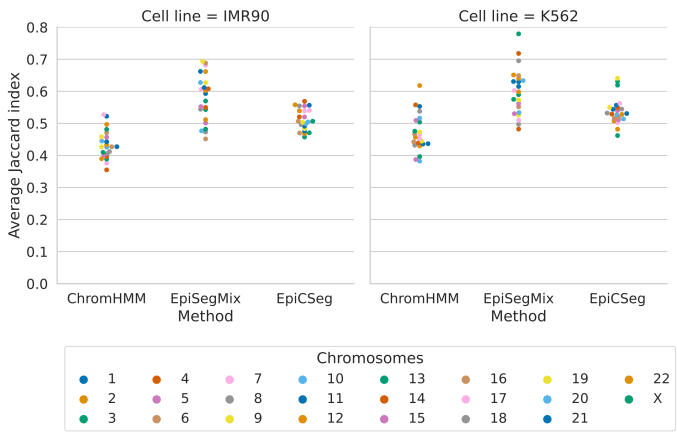


Figure 5: Robustness to replicate datasets: All three methods were fitted on each chromosome separately on the replicate datasets of IMR90 and K562 using an 8-state HMM. Each point represents the Jaccard similarity between two segmentations for a distinct chromosome, with a higher Jaccard index indicating a greater overlap.

to model state durations that follow a negative binomial instead of a geometric distribution, which better reflects the inherent segment length of chromatin states that cover either small peaks or broad domains.

A comparison with ChromHMM and EpiCSeg proposes that EpiSegMix has the potential to provide more accurate and robust segmentations. In particular, enabling a more accurate description of chromatin states linked to constitutive and Polycomb associated heterochromatin. The influence of the modified HMM topology on the segmentation suggests that testing other topology models is an important aspect to increase the modeling accuracy. Moreover, reducing the noise in ChIP-seq data are important considerations for future work.

In summary, we showed that the modeling assumption of the HMM have an impact on the segmentation quality and biological interpretation. Due to its high flexibility, EpiSegMix allows to accurately fit an HMM to read count data with varying distributional properties and provides the additional option of a more flexible duration modeling. With EpiSegMix, we therefore provide a configurable framework for chromatin segmentation that can be applied to a wide range of data.

6 Author contributions

The authors wish it to be known that J.S. and N.A. contributed equally to this work. The method was developed by J.S. and N.A. under the supervision of J.W., A.S. and S.R.. L.L. performed preliminary analyses to select suitable distributions. J.S, N.A. and S.R. contributed to writing and all authors contributed to revising the manuscript.

7 Ethics declaration

None declared.

8 Conflict of interests

None declared.

9 Funding

Internal funding.

10 Data availability

The data can be downloaded from the ENCODE portal. All accession numbers are provided in Supplementary Table 2.

11 Acknowledgments

We acknowledge the ENCODE consortium for producing the data.

References

Allis, C. D. and Jenuwein, T. (2016). The molecular hallmarks of epigenetic control. *Nature Reviews Genetics*, 17(8):487–500. Number: 8 Publisher: Nature Publishing Group.

Baker, M. (2011). Making sense of chromatin states. *Nature Methods*, 8(9):717–722. Number: 9 Publisher: Nature Publishing Group.

Bannister, A. J. and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Research*, 21(3):381–395. Number: 3 Publisher: Nature Publishing Group.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, 129(4):823–837.

Beacon, T. H., Delcuve, G. P., López, C., Nardocci, G., Kovalchuk, I., van Wijnen, A. J., and Davie, J. R. (2021). The dynamic broad epigenetic (H3K4me3, H3K27ac) domain as a mark of essential genes. *Clinical Epigenetics*, 13(1):138.

Bernoulli, J. (1713). *Ars Conjectandi*.

Bilmes, J. (1998). A gentle tutorial of the em algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *CTIT technical reports series*.

Blackledge, N. P. and Klose, R. J. (2021). The molecular principles of gene regulation by Polycomb repressive complexes. *Nature Reviews Molecular Cell Biology*, 22(12):815–833. Number: 12 Publisher: Nature Publishing Group.

Brun, R. and Rademakers, F. (1997). ROOT — An object oriented data analysis framework. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 389(1):81–86.

Bujold, D., Morais, D. A. d. L., Gauthier, C., et al. (2016). The International Human Epigenome Consortium Data Portal. *Cell Systems*, 3(5):496–499.e2. Publisher: Elsevier.

Carter, B. and Zhao, K. (2021). The epigenetic basis of cellular heterogeneity. *Nature Reviews Genetics*, 22(4):235–250. Number: 4 Publisher: Nature Publishing Group.

Cavalli, G. and Heard, E. (2019). Advances in epigenetics link genetics to the environment and disease. *Nature*, 571(7766):489–499. Number: 7766 Publisher: Nature Publishing Group.

Chan, R. C. W., Libbrecht, M. W., Roberts, E. G., Bilmes, J. A., Noble, W. S., and Hoffman, M. M. (2018). Segway 2.0: Gaussian mixture models and minibatch training. *Bioinformatics*, 34(4):669–671.

Daneshpajouh, H., Chen, B., Shokraneh, N., Masoumi, S., Wiese, K. C., and Libbrecht, M. W. (2022). Continuous chromatin state feature annotation of the human epigenome. *Bioinformatics*, 38(11):3029–3036.

Day, N., Hemmaplardh, A., Thurman, R. E., Stamatoyannopoulos, J. A., and Noble, W. S. (2007). Unsupervised segmentation of continuous genomic data. *Bioinformatics*, 23(11):1424–1426.

Diaz, A., Park, K., Lim, D. A., and Song, J. S. (2012). Normalization, bias correction, and peak calling for ChIP-seq. *Statistical Applications in Genetics and Molecular Biology*, 11(3). Publisher: De Gruyter.

Dunham, I., Kundaje, A., Aldred, S. F., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74. Number: 7414, Publisher: Nature Publishing Group.

Ernst, J. and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, 28(8):817–825. Number: 8 Publisher: Nature Publishing Group.

Ernst, J. and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods*, 9(3):215–216. Number: 3 Publisher: Nature Publishing Group.

Fisher, R. A. (1941). The negative binomial distribution. *Annals of Eugenics*, 11(1):182–187.

Griffiths, D. A. (1973). Maximum Likelihood Estimation for the Beta-Binomial Distribution and an Application to the Household Distribution of the Total Number of Cases of a Disease. *Biometrics*, 29(4):637–648.

Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. A., and Noble, W. S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods*, 9(5):473–476. Number: 5 Publisher: Nature Publishing Group.

Johnson, N. L., Kotz, S., Kemp, A. W., et al. (1993). *Univariate discrete distributions*. Wiley, New York, NY [u.a.], 2nd ed edition.

Lee, K.-E. and Park, H.-S. (2014). A Review of Three Different Studies on Hidden Markov Models for Epigenetic Problems: A Computational Perspective. *Genomics & Informatics*, 12(4):145–150.

Libbrecht, M. W., Chan, R. C. W., and Hoffman, M. M. (2021). Segmentation and genome annotation algorithms for identifying chromatin state and other genomic patterns. *PLOS Computational Biology*, 17(10):e1009423. Publisher: Public Library of Science.

Lindén, A. and Mäntyniemi, S. (2011). Using the negative binomial distribution to model overdispersion in ecological count data. *Ecology*, 92(7):1414–1421.

Mammana, A. and Chung, H.-R. (2015). Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome. *Genome Biology*, 16(1):151.

Mammana, A. and Helmuth, J. (2023). bamsignals: Extract read count signals from bam files.

Mann, T. P. (2006). Numerically Stable Hidden Markov Model Implementation. `https://citeseerx.ist.psu.edu/doc/10.1.1.322.9645`.

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., and Köster, J. (2021). Sustainable data analysis with Snakemake. Technical Report 10:33, F1000Research.

Park, P. J. (2009). ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680. Number: 10 Publisher: Nature Publishing Group.

Poisson, S. (1837). *Recherches sur la probabilité des jugements en matière criminelle et en matière civile: précédées des règles générales du calcul des probabilités*. Bachelier Paris.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286. Conference Name: Proceedings of the IEEE.

Rigby, R. A., Stasinopoulos, M. D., Heller, G. Z., and Bastiani, F. D. (2019). *Distributions for Modeling Location, Scale, and Shape: Using GAMLSS in R*. Chapman and Hall/CRC, New York.

Rodríguez-Avi, J., Conde-Sánchez, A., Sáez-Castillo, A. J., and Olmo-Jiménez, M. J. (2007). A new generalization of the Waring distribution. *Computational Statistics & Data Analysis*, 51(12):6138–6150.

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

Schröder, C. and Rahmann, S. (2017). A hybrid parameter estimation algorithm for beta mixtures and applications to methylation state classification. *Algorithms for Molecular Biology*, 12(1):21.

Sichel, H. S. (1992). Anatomy of the generalized inverse Gaussian-poisson distribution with special applications to bibliometric studies. *Information Processing & Management*, 28(1):5–17.

Stein, G. Z., Zucchini, W., and Juritz, J. M. (1987). Parameter estimation for the Sichel distribution and its multivariate extension. *Journal of the American Statistical Association*, 82(399):938–944.

Wang, Z. (2011). One mixed negative binomial distribution with application. *Journal of Statistical Planning and Inference*, 141(3):1153–1160.

Supplementary Methods

Notation for Hidden Markov Models (HMMs)

- observation $O = (O_1, \ldots, O_T)$ with T time points, where T is the number of rows in the count matrix
- finite set of hidden states $\{1, 2, \ldots, N\}$
- probability vector $(\pi_i)_{i \in \{1,2,\ldots,N\}}$, which defines the starting probabilities for all states
- $N \times N$ matrix A, where each entry a_{ij} denotes the transition probability to move from state i to state j
- complete collection of emission parameters $B = \{b_j(\cdot) \mid j \in \{1, 2, \ldots, N\}\}$
- probability distribution $P_\theta(\cdot)$, governed by a set of parameters θ

Parameter Estimation

The parameter estimation for an HMM with parameters $\theta = (\pi, A, B)$ is performed using the Baum-Welch algorithm (Rabiner, 1989). In each iteration r, the complete log likelihood function is maximized until convergence to a local optimum or until the maximum number of iterations has been reached. In our flexible distribution HMM, the complete log-likelihood function is given by Equation (6) (see also Mammana and Chung (2015, Supplementary Material))

$$\theta^{(r+1)} = \arg\max_\theta \left(\sum_{i=1}^{N} \gamma_i(1) \log \pi_i + \sum_{t=2}^{T} \sum_{i=1}^{N} \sum_{j=1}^{N} \xi_{ij}(t) \log a_{ij} + \sum_{t=1}^{T} \sum_{i=1}^{N} \gamma_i(t) \log P_{\theta_i}^{(i)}(O_t = o_t) \right), \tag{6}$$

where

$$\gamma_i(t) = P_\theta(Q_t = i \mid O), \qquad i \in \{1, \ldots, N\}, t \in \{1, \ldots, T\}, \tag{7}$$

$$\xi_{ij}(t) = P_\theta(Q_t = i, Q_{t+1} = j \mid O), \qquad i, j \in \{1, \ldots, N\}, t \in \{1, \ldots, T\}, \tag{8}$$

which are efficiently computed using the forward-backward algorithm (Rabiner, 1989), resulting in the following update rules (Bilmes, 1998):

$$\pi_i^{(r+1)} = \gamma_i(1) \tag{9}$$

$$a_{ij}^{(r+1)} = \frac{\sum_{t=1}^{T-1} \xi_{ij}(t)}{\sum_{t=1}^{T-1} \gamma_i(t)} \tag{10}$$

$$\theta_i^{(r+1)} = \arg\max_{\theta_i} \sum_{t=1}^{T} \gamma_i(t) \log P_{\theta_i}^{(i)}(o_t) \tag{11}$$

For an observation with M histone marks, the update formula for the emission parameters in state i is given by

$$\arg\max_{\theta_i} \sum_{t=1}^{T} \gamma_i(t) \log P_{\theta_i}^{(i)}(o_t) = \arg\max_{\theta_i} \sum_{t=1}^{T} \gamma_i(t) \log \left(\prod_{m=1}^{M} P_{\theta_{im}}^{(im)}(o_{tm}) \right) \tag{12}$$

$$= \arg\max_{\theta_i} \sum_{m=1}^{M} \sum_{t=1}^{T} \gamma_i(t) \log P_{\theta_{im}}^{(im)}(o_{tm}) \tag{13}$$

Thus, θ can be optimized independently for each state i and each histone mark m:

$$\theta_{im}^{(r+1)} = \arg\max_{\theta_{im}} \sum_{t=1}^{T} \gamma_i(t) \log P_{\theta_{im}}^{(im)}(o_{tm}) \tag{14}$$

Since for most distributions no closed-form solution exists for Equation (14), it is optimized numerically using the *MIGRAD()* function from the *ROOT Minuit 2* package (Brun and Rademakers, 1997).

In the extended-state topology HMM, for each state i, we create a sub-HMM with states $i_1, i_2, \ldots, i_{S_i}$, where S_i is the number of states in the sub-HMM of state i. The update formulas enforcing the additional constraints of equal emission and self-transition probabilities are given by

$$\theta_{im}^{(r+1)} = \arg\max_{\theta_{im}} \sum_{t=1}^{T} \left(\sum_{s=1}^{S_i} \gamma_{i_s}(t) \right) \log P_{\theta_{im}}^{(im)}(o_{tm}) \tag{15}$$

$$a_{ii}^{(r+1)} = \frac{\sum_{t=1}^{T-1} \sum_{s=1}^{S_i} \xi_{i_s i_s}(t)}{\sum_{t=1}^{T-1} \sum_{s=1}^{S_i} \gamma_{i_s}(t)} \tag{16}$$

All other parts of the Baum–Welch algorithm remain unchanged.

Emission Modeling

We subsequently provide the probability mass function, the method-of-moments estimators for initialization and if available the closed form solution for maximum likelihood estimation (MLE) for all implemented discrete distributions.

For the method-of-moments estimators, the sample mean $\bar{x}$ and sample variance s^2 are given for a sample $(x_1, \ldots, x_n)$ by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{17}$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{18}$$

Binomial Distribution

The probability mass function of a binomial distribution (BI) with parameters $0 \leq p \leq 1$ and a natural number $n > 0$ is given by (Bernoulli, 1713)

$$P(X = k) = \binom{n}{k} p^k \, (1-p)^{n-k} \quad \text{for } k = 0, 1, 2, \ldots, n \,, \tag{19}$$

where $\binom{n}{k}$ is the binomial coefficient defined as $\frac{n!}{k!(n-k)!}$. Assuming that n is known, the method-of-moments estimator of p is given by

$$\hat{p} = \frac{\bar{x}}{n} \tag{20}$$

The MLE of p in state i is given for an observation $O = (o_1, \ldots, o_t)$ and membership coefficients $\gamma_i(t)$ by

$$\hat{p}_i = \frac{\sum_{t=1}^{T} \gamma_i(t) \cdot o_t}{n \sum_{t=1}^{T} \gamma_i(t)} \tag{21}$$

If n and p are unknown, the estimator of n is not guaranteed to be finite. Thus, we use the sample maximum as an estimator of n, as proposed by Fisher (1941).

Poisson Distribution

The probability mass function of a Poisson distribution (PO) with parameter $\lambda > 0$ is given by (Poisson, 1837)

$$P(X = k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!} \text{ for } k = 0, 1, 2, \ldots \,. \tag{22}$$

The method-of-moments estimator for λ is given by the sample mean (Johnson et al., 1993, ch. 4), and the MLE is computed for an observation $O = (o_1, \ldots, o_t)$ and membership coefficients $\gamma_i(t)$ by

$$\hat{\lambda}_i = \frac{\sum_{t=1}^{T} \gamma_i(t) \cdot o_t}{\sum_{t=1}^{T} \gamma_i(t)} \tag{23}$$

Negative Binomial Distribution

The probability mass function of a negative binomial distribution (NBI) with parameters $r \geq 0$ and $0 \leq p \leq 1$ is given by

$$P(X = k) = \binom{k+r-1}{r-1} (1-p)^k \, p^r \quad \text{for } k = 0, 1, 2, \ldots \,. \tag{24}$$

The method-of-moments estimators are given by (Lindén and Mäntyniemi, 2011)

$$\hat{r} = \frac{\bar{x}^2}{s^2 - \bar{x}} \,, \tag{25}$$

$$\hat{p} = \frac{\bar{x}}{s^2} \,. \tag{26}$$

Beta Binomial Distribution

The probability mass function of a beta binomial distribution (BB) with parameters $\alpha > 0$, $\beta > 0$ and a natural number $n > 0$ is given by (Griffiths, 1973)

$$P(X = k) = \binom{n}{k} \frac{B(\alpha + k, n + \beta - k)}{B(\alpha, \beta)} \quad \text{for } k = 0, 1, \ldots, n\,, \tag{27}$$

where

$$B(\alpha, \beta) = \frac{\Gamma(\alpha) \cdot \Gamma(\beta)}{\Gamma(\alpha + \beta)} \tag{28}$$

is the Beta function and

$$\Gamma(x) = \int_0^\infty t^{x-1} \mathrm{e}^{-t}\, \mathrm{d}t \quad \text{for } x > 0 \tag{29}$$

is the the Gamma function (Johnson et al., 1993, p.4). Other names for the beta binomial distribution are Pólya or negative hypergeometric distribution. The method-of-moments estimators of α and β are given by

$$\alpha = \frac{\pi}{\theta}\,, \tag{30}$$

$$\beta = \frac{1 - \pi}{\theta}\,, \tag{31}$$

where π and θ are estimated by

$$\hat{\pi} = \frac{\bar{x}}{n}\,, \tag{32}$$

$$\hat{\theta} = \frac{\hat{s}^2 - n\hat{\pi}(1 - \hat{\pi})}{n^2\hat{\pi}(1 - \hat{\pi}) - \hat{s}^2}\,. \tag{33}$$

Similar to the binomial distribution, if n is unknown, it is estimated by the sample maximum.

Beta Negative Binomial Distribution

The probability mass function of a beta negative binomial distribution (BNB) with parameters $r > 0$, $\alpha > 0$, $\beta > 0$ is given by (Johnson et al., 1993, ch. 6.2.3)

$$P(X = k) = \frac{\Gamma(k + r)}{\Gamma(k + 1)\Gamma(r)} \frac{B(\alpha + r, \beta + k)}{B(\alpha, \beta)} \quad \text{for } k = 0, 1, \ldots\,, \tag{34}$$

where $B(\cdot, \cdot)$ denotes the Beta function (28) and $\Gamma(\cdot)$ denotes the Gamma function (29). Other names for the beta negative binomial distribution are generalized Waring distribution, inverse Markov–Pólya distribution or beta–Pascal distribution.

Since the n–th theoretical moment is only finite if $\alpha > n$ and in addition r and β are interchangeable (Rodríguez-Avi et al., 2007), a simplified approach is used to calculate the method-of-moment estimators. First, r and p are estimated under the assumption of a negative binomial distribution. The method-of-moments estimators of α and β are then estimated for a beta distribution with the sample mean given by p and the sample variance given by an arbitrarily chosen value of 0.05. For $\bar{x} = p$ and $s^2 = 0.05$, α and β are then given by (Schröder and Rahmann, 2017)

$$\psi = \frac{\bar{x} \cdot (1 - \bar{x})}{s^2} - 1 \tag{35}$$

$$\alpha = \bar{x} \cdot \psi \tag{36}$$

$$\beta = (1 - \bar{x}) \cdot \psi \tag{37}$$

Since the method-of-moments estimators are only used for initialization of the Baum-Welch algorithm, inaccurate estimates are acceptable.

Sichel Distribution

The probability mass function of a Sichel distribution with parameters $\mu > 0$, $\sigma > 0$ and $-\infty < v < \infty$ is given by (Sichel, 1992)

$$P(X = k) = \frac{\mu^k\, K_{k+v}(\alpha)}{k!\, (\alpha \cdot \sigma)^{k+v} K_v\left(\frac{1}{\sigma}\right)} \quad \text{for } k = 0, 1, 2 \ldots\,, \tag{38}$$

where $\alpha = \sqrt{\frac{1}{\sigma^2} + \frac{2\mu}{\sigma}}$ and $K_v(t)$ denotes the modified Bessel function of the second kind of order v and argument t given by

$$K_v(t) = \frac{1}{2} \int_0^\infty x^{v-1} \mathrm{e}^{-\frac{1}{2}t(x+x^{-1})} \, \mathrm{d}x \,. \tag{39}$$

To avoid numerical problems during the computation of the Bessel function for large orders, the probability mass function is calculated recurrently based on the following recurrence relation (Stein et al., 1987)

$$P(X=0) = \left(\frac{w}{\alpha}\right)^v \frac{K_v(\alpha)}{K_v(w)} \,, \tag{40}$$

$$P(X=1) = P(X=0) \cdot \frac{\mu \cdot w}{\alpha} \, \frac{K_{v+1}(\alpha)}{K_v(\alpha)} \,, \tag{41}$$

$$P(X=k) = \frac{2\mu w}{\alpha^2} \cdot \frac{k+v-1}{k} \cdot P(X=k-1) + \frac{(\mu w/\alpha)^2}{k(k-1)} \cdot P(X=k-2) \tag{42}$$

with $w = \sqrt{\mu^2 + \alpha^2} - \mu$.

For $v = -\frac{1}{2}$ the Sichel distribution is equal to the inverse Gaussian Poisson distribution (Stein et al., 1987). For initialization, the parameters of the Sichel distribution are estimated using the method-of-moments estimators of the inverse Gaussian Poisson distribution, given by

$$v = -\frac{1}{2} \,, \tag{43}$$

$$\mu = \bar{x} \,, \tag{44}$$

$$\sigma = \frac{\sqrt{\alpha^2 + \bar{x}^2} + \bar{x}}{\alpha^2} \,, \tag{45}$$

with $w = \frac{\bar{x}}{s^2/\bar{x} - 1}$ and $\alpha^2 = (w+\mu)^2 - \mu^2$.

Zero-Adjusted Distributions

A zero-adjusted distribution is a discrete mixture distribution that has a fixed zero probability π. We currently support the zero-adjusted Poisson, negative binomial, beta negative binomial and Sichel distribution. For a discrete random variable X', the zero adjusted probability mass function with zero probability $0 \leq \pi \leq 1$ is given by (Rigby et al., 2019)

$$P(X=k) = \begin{cases} \pi & k = 0 \\ (1-\pi) \cdot \frac{P(X'=k)}{1-P(X'=0)} & k \neq 0 \end{cases} \tag{46}$$

The method-of-moments estimator of π is given by the zero frequency and the MLE of π is given by

$$\hat{\pi}_i = \frac{\sum_{t=1}^T \gamma_i(t) \cdot \mathbb{1}_{o_t=0}}{\sum_{t=1}^T \gamma_i(t)} \tag{47}$$

where $\gamma_i(t)$ denotes the membership coefficient of observation o_t in state i. The parameters of the distribution for X' can be optimized independently.

Supplementary Figures and Tables

Selecting the Number of States

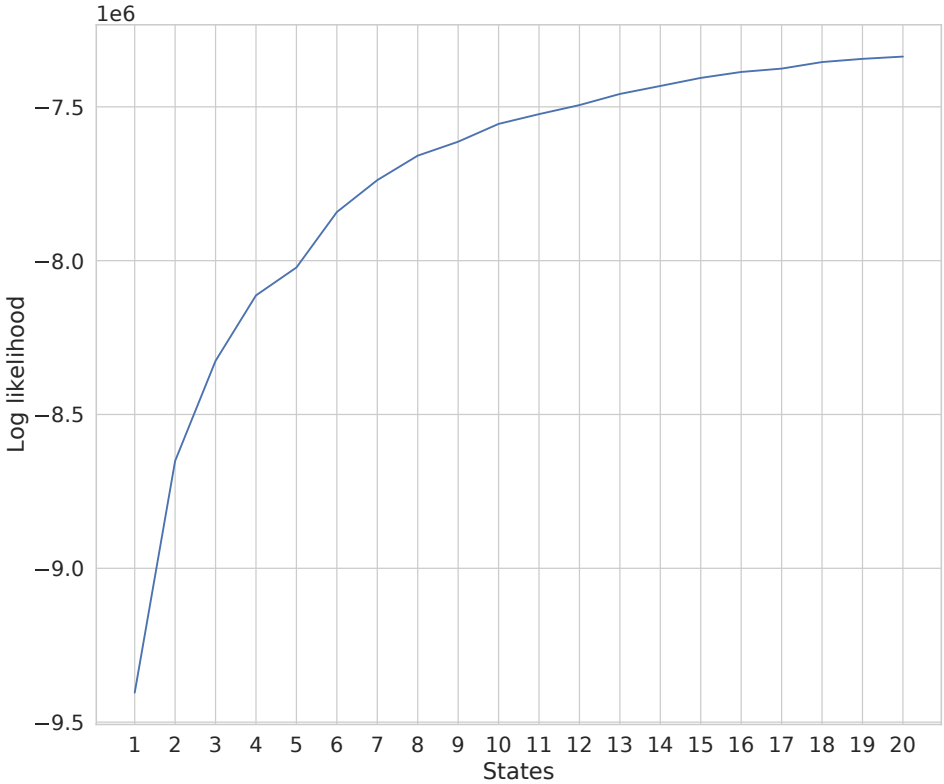


Figure 6: Log-likelihoods for a 1-state HMM up to a 20-state HMM for the IMR90_1 dataset using the six core histone marks H3K9me3, H3K27me3, H3K36me3, H3K4me1, H3K4me3 and H3K27ac. A higher log likelihood indicates a better model fit. Since the log-likelihood of an HMM increases monotonically as the number of states increases, the number of states that we choose should be a good compromise between model fit and interpretability.

Selecting the Most Appropriate Distributions

Table 2: Selected distribution for each mark per dataset. To determine the best-suited distributions, we fitted a 3-state HMM for each mark individually and we chose the distribution with the highest log-likelihood for the combined model. NBI denotes the negative binomial, SI the Sichel and BNB the beta negative binomial distribution.

	IMR90_1	IMR90_2	H1_1	K562_1	K562_2
H3K27me3	SI	BNB	SI	SI	BNB
H3K9me3	NBI	NBI	SI	NBI	SI
H3K36me3	SI	SI	BNB	NBI	NBI
H3K4me3	SI	SI	SI	SI	SI
H3K4me1	SI	SI	NBI	NBI	NBI
H3K27ac	BNB	SI	SI	BNB	SI

Segmentation Comparison for IMR90

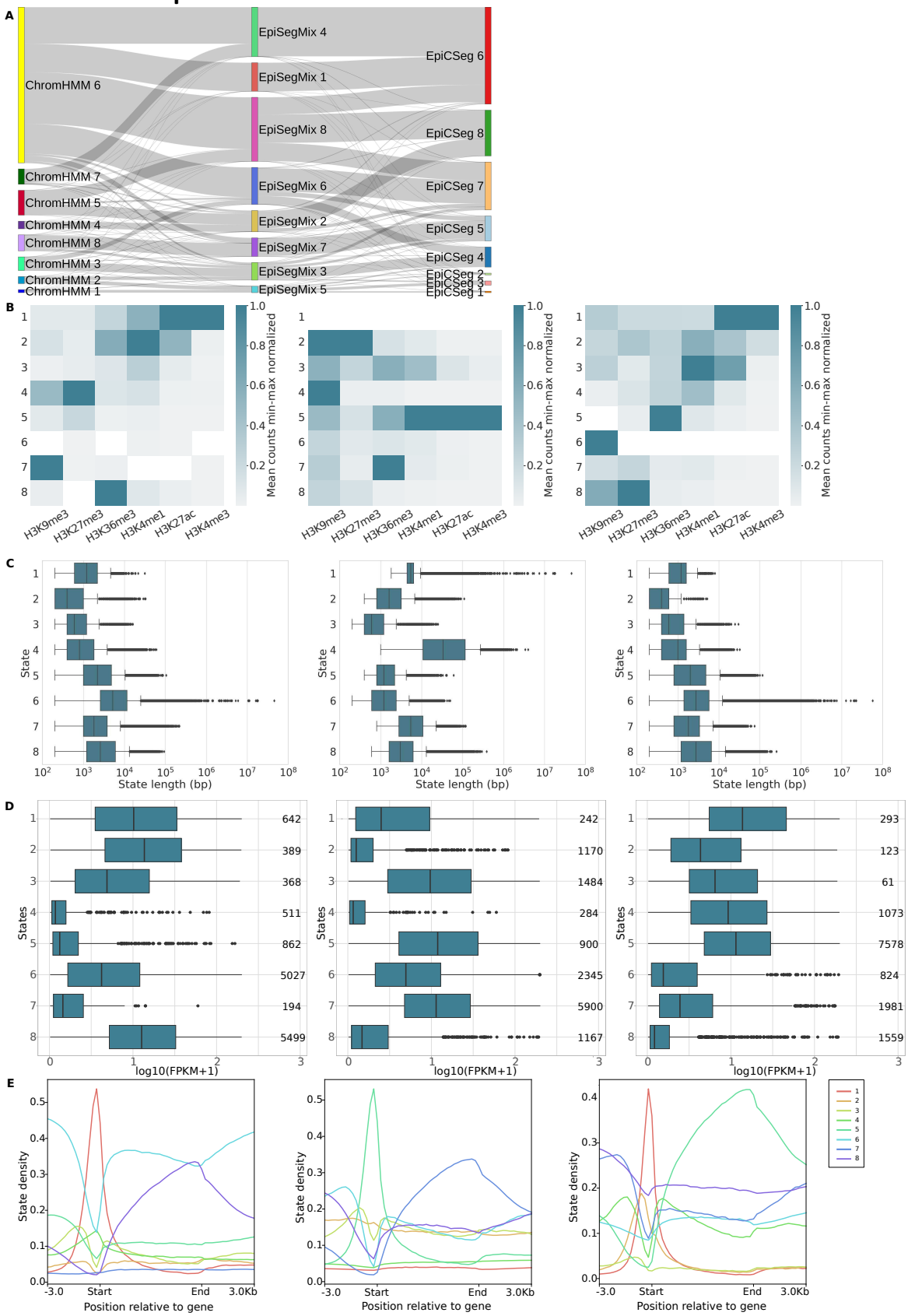


Figure 7: Comparison between ChromHMM (left), EpiSegMix (middle) and EpiCSeg (right) for IMR90_1. **A** State overlap between methods. The bar widths corresponds to the genome coverage of each state and the edge widths to the overlap between segmentations. **B** Normalized histone enrichment per state. **C** Segment length per state. **D** Gene expression per state. **E** State distribution around transcribed genes.

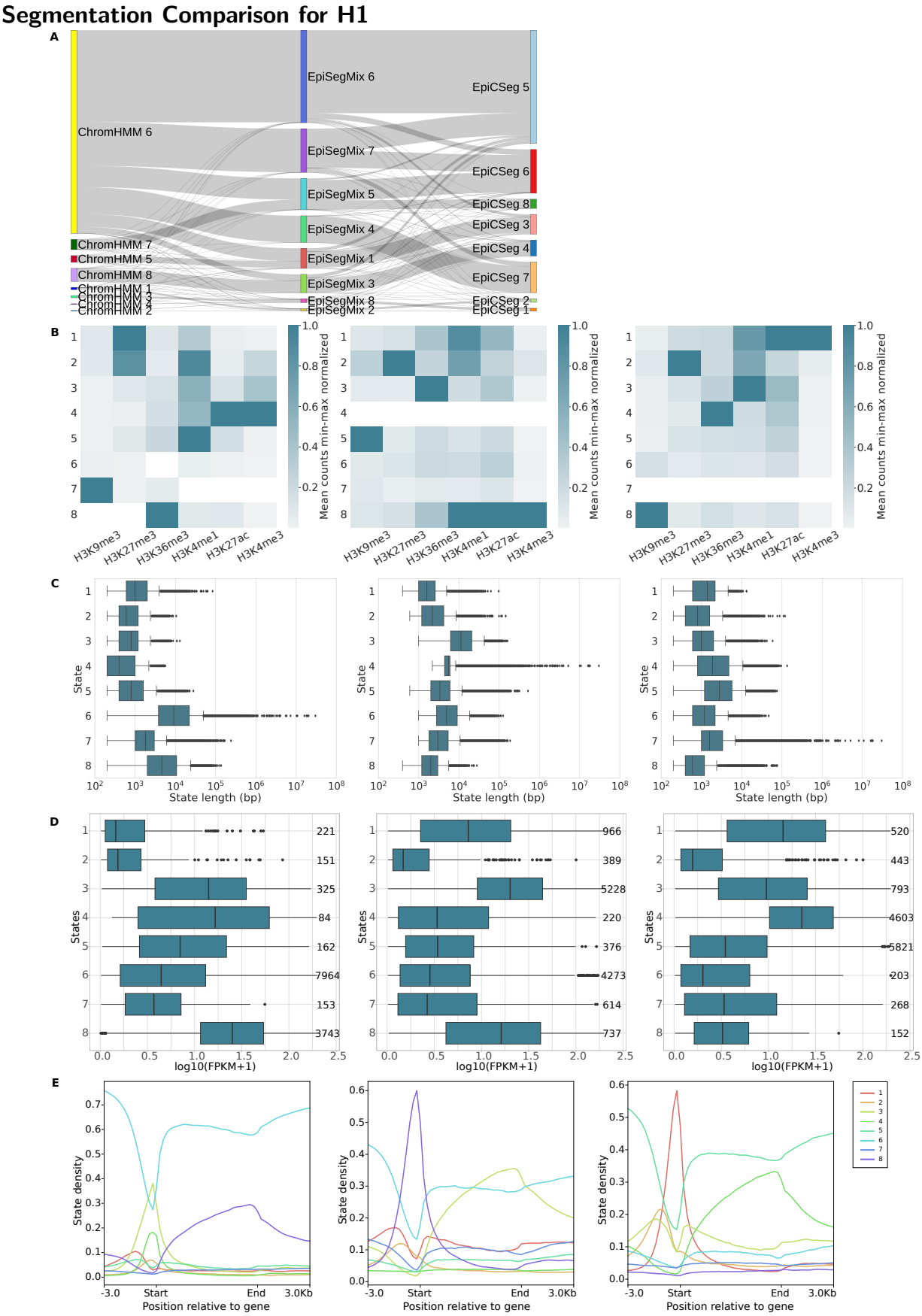


Figure 8: Comparison between ChromHMM (left), EpiSegMix (middle) and EpiCSeg (right) for H1_1. **A** State overlap between methods. The bar widths corresponds to the genome coverage of each state and the edge widths to the overlap between segmentations. **B** Normalized histone enrichment per state. **C** Segment length per state. **D** Gene expression per state. **E** State distribution around transcribed genes.

Data Availability

Table 3: Accession numbers of all files downloaded from ENCODE.

	IMR90_1	IMR90_2	H1_1
H3K9me3	ENCFF648MBH	ENCFF407SIR	ENCFF597CRW
H3K27me3	ENCFF825OWT	ENCFF868WTE	ENCFF596SHE
H3K36me3	ENCFF538AFE	ENCFF433YUN	ENCFF295LHK
H3K4me1	ENCFF969ZSU	ENCFF351EKD	ENCFF322LHI
H3K4me3	ENCFF501ZHJ	ENCFF175EZW	ENCFF340UJK
H3K27ac	ENCFF146UYU	ENCFF071VOI	ENCFF238SQN
RNA-seq	ENCFF833OTW	ENCFF496RIW	ENCFF723ONV

	K562_1	K562_2
H3K9me3	ENCFF146NLP	ENCFF559DHZ
H3K27me3	ENCFF190OWE	ENCFF692KQZ
H3K36me3	ENCFF639PLN	ENCFF673KBG
H3K4me1	ENCFF778EZR	ENCFF580LGK
H3K4me3	ENCFF236SNL	ENCFF661UGK
H3K27ac	ENCFF301TVL	ENCFF879BWC
RNA-seq	ENCFF104VTJ	ENCFF201HGA