JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 27, Number 0, 2020
© Mary Ann Liebert, Inc.
Pp. 1–16
DOI: 10.1089/cmb.2019.0280

Research Article

ModHMM: A Modular Supra-Bayesian Genome Segmentation Method

PHILIPP BENNER and MARTIN VINGRON

ABSTRACT

Genome segmentation methods are powerful tools to obtain cell type or tissue-specific genome-wide annotations and are frequently used to discover regulatory elements. However, traditional segmentation methods show low predictive accuracy and their data-driven annotations have some undesirable properties. As an alternative, we developed ModHMM, a highly modular genome segmentation method. Inspired by the supra-Bayesian approach, it incorporates predictions from a set of classifiers. This allows to compute genome segmentations by utilizing state-of-the-art methodology. We demonstrate the method on ENCODE data and show that it outperforms traditional segmentation methods not only in terms of predictive performance, but also in qualitative aspects. Therefore, ModHMM is a valuable alternative to study the epigenetic and regulatory landscape across and within cell types or tissues.

Keywords: genome segmentation, HMM, supra-Bayesian.

1. INTRODUCTION

A single organism may consist of a remarkable diversity of cell types all sharing the same genotype. To understand how this diversity arises, current research in molecular biology has focused much attention on the functioning of transcriptional regulation. Genome-wide measurements of epigenetic marks and RNA expression have recently become available for many cell types and tissues (ENCODE Project Consortium, 2012). These data provide a first glimpse at the regulatory program on a genome-wide scale. It is used to annotate regulatory elements and to locate important switches that control cell identity (Hoffman et al., 2012b).

Genome segmentations are frequently used as a starting point for the identification and analysis of regulatory elements within cell types or tissues. By combining data from multiple experiments a genome segmentation method assigns a chromatin state to each genomic position. This may include regulatory elements such as active or repressed promoters and enhancers, active transcription or regions without an apparent function. The set of chromatin states a segmentation method is able to detect heavily depends on the choice of features. Typically a variety of histone modification ChIP-seq experiments is used possibly in combination with measurements of chromatin accessibility. However, other sources of information may be used as well, including DNA methylation, CpG content, or evolutionary conservation. Chromatin states are characterized by specific combinations of such features. For instance, promoters are often conserved

Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany.

elements and accessible for transcription factors where the flanking nucleosomes are marked by H3K4me3. In contrast, most enhancers are less conserved and marked by H3K4me1. Enrichment of H3K27ac is found at active promoters and enhancers, whereas H3K27me3 is known to be a repressive mark observed at bivalent promoters and poised enhancers (Barski et al., 2007; Heintzman et al., 2007; Calo and Wysocka, 2013; Heinz et al., 2015).

Most common segmentation methods are instances of hidden Markov models (HMMs) where the observed data are assumed to be caused by an unobserved sequence of hidden states with Markov dependency structure. A frequently used implementation of this type is ChromHMM (Ernst and Kellis, 2012), which, however, relies on binarized data. A more advanced HMM-based method is EpiCSeg (Mammana and Chung, 2015) that addresses this shortcoming by using negative multinomial distributions to model observations. The handling of both methods is seemingly easy. Parameters are estimated unsupervisedly without the need for a training set using a maximum likelihood approach. Afterward, by inspection of estimated parameters each hidden state is identified with one or more chromatin states. Although this approach is very easy to apply, it also bears several risks. The specific combination of features known to mark a chromatin state and their spatial distribution is often not well reflected by the model.

Hence, supervised methods specialized in the detection of regulatory elements typically perform much better. To obtain good classification performances of unsupervised HMMs the number of hidden states often exceeds the actual observed number of chromatin states. This leads to highly fragmented genome segmentations where single chromatin states are represented by multiple states of the HMM. Figure 1 illustrates this problem on a promoter of a transcribed gene. An optimal segmentation would detect the region as a single active promoter with a transcribed region to the left. However, typical segmentations obtained with ChromHMM or EpiCSeg instead show a highly fragmented promoter region. Another drawback of the unsupervised HMM approach is the low flexibility of the model offering no glaring way to improve a poor segmentation. ChromHMM and EpiCSeg have only two parameters, namely the number of hidden states and the genomic bin size, whose effect on the resulting segmentation is highly unpredictable. Furthermore, to determine the optimal set of parameters it is necessary to learn and evaluate a large number of different models, effectively negating the presumed simplicity.

ChroModule (Won et al., 2013) is a supervised alternative that models the spatial distribution of features at chromatin states with left–right structured HMMs that are commonly used in speech recognition (Rabiner, 1989). However, the construction of a model requires a training set for each chromatin state. As an alternative to the HMM-based methods, Segway (Hoffman et al., 2012a,b) allows to compute segmentations based on arbitrary hidden processes, as long as the model can be represented as a dynamic Bayesian network. It operates on a single base-pair resolution and with its default model computes segmentations that are even more fine-grained than those of ChromHMM and EpiCSeg (Hoffman et al., 2012a). Segway models are typically trained on a small fraction of the available data, due to the computational complexity of the inference algorithm and the high data resolution.

A different approach is implemented in StateHub-StatePaintR (Coetzee et al., 2018). Instead of inferring chromatin states de novo every time a segmentation is computed, the method implements a model of our current understanding of chromatin state characteristics. In particular, the model encodes for each chromatin state the set of features that are positively or negatively associated with that state. The method is certainly a step toward the right direction, but it also does not model the spatial distribution of features around chromatin states and, therefore, produces highly fragmented segmentations, similar to those of unsupervised HMMs.

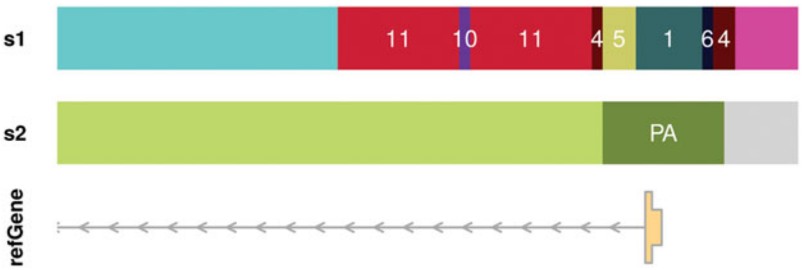

FIG. 1. Genome segmentations. Typical genome segmentation (s1) where the promoter is fragmented into many different segments. In this example, the optimal segmentation (s2) shows a single active promoter segment (PA) with a transcribed region to the left and no signal to the right.

We develop in this study a new modular segmentation method based on HMMs called ModHMM that addresses some of these shortcomings in the following way. First of all, we recognize that jointly learning all parameters of an HMM in an unsupervised way is overly ambitious and leads to poor results. Instead, we assemble the segmentation method piece by piece allowing us to guide the learning process as much as possible. Second, our objective is to construct a method that may benefit from the ample variety of well-performing classifiers that have been developed for most regulatory elements. Inspired by the supra-Bayesian approach (Lindley et al., 1979; Lindley, 1982, 1985; Genest and Zidek, 1986; Gelfand et al., 1995; Jacobs, 1995), we construct an HMM that acts as a decision maker who integrates assessments from several experts. Each expert or classifier is specialized in the detection of a single chromatin state, possibly by considering only a subset of the available features. The classifiers may also model the spatial distribution of features near functional elements to improve prediction accuracy.

Hence, our segmentation method consists of an HMM combined with a set of chromatin state classifiers. As opposed to traditional segmentation methods, our HMM does not take feature tracks (i.e., ChIP-seq/ATAC-seq tracks) as input, but instead regards the genome-wide probability assessments of the chromatin state classifiers as observations. We constructed the method in a highly modular way, allowing to easily improve segmentations by replacing single classifiers. To facilitate the usage of ModHMM, we constructed a default set of chromatin state classifiers. The parameters of supervised classifiers are commonly estimated on some training set. However, constructing a training set for each classifier would not only be a tedious task, but it would also shift control over the resulting segmentation to the composition of such training sets. Instead, we do not rely on a training set but engineer each classifier by translating contemporary knowledge of chromatin states into a probabilistic model. This makes ModHMM easily applicable to new data sets and ensures that the interpretation of hidden states is consistent across samples.

We evaluate ModHMM equipped with its default chromatin state classifiers on promoter and enhancer test sets and show that it outperforms traditional segmentation methods not only in classification accuracy but also in qualitative aspects, meaning that ModHMM segmentations are less fragmented. To improve segmentation results, classifiers from the default set can be easily replaced by more powerful supervised alternatives, allowing to incorporate predictions from state-of-the-art methods. Inspired by the DFilter peak calling method (Kumar et al., 2013), we develop a classifier that models the spatial distribution of features around chromatin states. We train this classifier on a set of active enhancers and show that it can be effectively integrated into our genome segmentation method.

2. METHODS

We consider chromatin states that are most relevant for the analysis of gene regulation and that are typically found in genome annotation studies (Hoffman et al., 2012b; Ernst and Kellis, 2017; Gorkin et al., 2017). This includes active promoters (PA) and enhancers (EA), primed (PR) and bivalent (BI) regions, as well as regions of active (TR) and low transcription (TL). In addition, we model heterochromatic regions marked by H3K27me3 (R1) or H3K9me3 (R2) and regions where either no signal (NS) or a control signal (CL) is observed. Enhancers and promoters are detected based on ATAC-seq (Buenrostro et al., 2013, 2015) data measuring chromatin accessibility in combination with histone marks H3K4me1 and H3K4me3. To measure the activity of promoters and enhancers we use histone mark H3K27ac (Creyghton et al., 2010). Although active promoters and enhancers can be accurately discriminated by the ratio of histone marks H3K4me1 and H3K4me3, we observed that the prediction accuracy is much lower for bivalent promoters and poised enhancers (Section 3.3).

Therefore, we decided to merge both chromatin states into a single bivalent state (BI), which is marked by H3K27me3 and H3K4me1 or H3K4me3. Similarly, we define primed states (PR) as accessible regions marked by H3K4me1 or H3K4me3 but showing no H3K27ac and H3K27me3 signal. We also model regions solely marked by either H3K27me3 (R1) or H3K9me3 (R2). Histone mark H3K27me3, catalyzed by the polycomb repressive complex 2, is involved in gene silencing and associated with constitutive heterochromatin (Kuzmichev et al., 2002; Margueron and Reinberg, 2011). In contrast, histone mark H3K9me3 is associated with constitutive heterochromatin, predominantly formed in gene-empty regions (Saksouk et al., 2015). Transcribed regions are typically marked by H3K36me3 (Wagner and Carpenter, 2012); however, we decided to use polyA RNA-seq data instead since it is a more direct and less noisy measurement. Our model also accounts for low levels of transcription (TL) that frequently occur at repressed genes or in intergenic regions, for instance near certain types of enhancers that generate unidirectional polyA+ eRNAs (Koch et al., 2011).

ModHMM is a highly modular genome segmentation method that incorporates predictions from a set of classifiers. In contrast to unsupervised methods, we manually construct most parts of the model. It consists of two components, the HMM and the set of chromatin state classifiers, both will be outlined in the following. For the chromatin state classifiers we consider a default set of engineered classifiers as well as a supervised alternative.

2.1. *Hidden Markov model*

ModHMM implements a HMM, which consists of an unobserved Markov process generating a series of hidden states, each emitting a single observation (Cappé et al., 2005). To define the HMM, we first must assign each chromatin state to one or more hidden states of the HMM and decide on a set of feasible transitions of the unobserved Markov process. Finally, we present the emission model for incorporating genome-wide predictions of the classifiers and show that transition rates must be further constrained to construct a well-functioning model.

2.1.1. Hidden states and feasible transitions. Unsupervised HMM-based segmentations methods, including ChromHMM and EpiCSeg, learn transition rates using a maximum likelihood approach and initially allow transitions between any two states. To guide the learning process, it is often helpful to enforce a predefined structure on the transition matrix (Galassi et al., 2007). In genetics such structured HMMs have been utilized before, for instance, for the prediction of gene structures from DNA sequences (Burge and Karlin, 1997). In our case, the structure of the HMM should encode any prior knowledge about the context in which chromatin states appear in the genome and may, for instance, be used to implement a model for actively transcribed genes. However, one has to be cautious not to enforce an overly simplistic model. For instance, an HMM that requires each gene to have exactly one promoter and a single transcribed region would not be realistic and result in wrong predictions. The converse, a model that is excessively complex would have equally poor predictive performance.

Therefore, we decided on an HMM with minimal structure, as depicted in Figure 2. Some chromatin states are represented by multiple hidden states of the HMM. For instance, active enhancers (EA) are represented by hidden states EA, EA$_1$, and EA$_2$ to model different contexts in which enhancers may appear. The HMM structure enforces that each transcribed region must be flanked by at least one active promoter and each active promoter must be flanked by a transcribed region. It also forbids that transcribed regions are flanked by active enhancers and primed or bivalent regions. In addition, transitions between active promoters, active enhancers, primed, and bivalent states are forbidden.

2.1.2. Emissions. The ModHMM segmentation method is inspired by the supra-Bayesian approach that integrates predictions of an expert committee from which a decision maker reaches a final decision. The expert committee consists of a set of classifiers, each specialized in the detection of a single chromatin

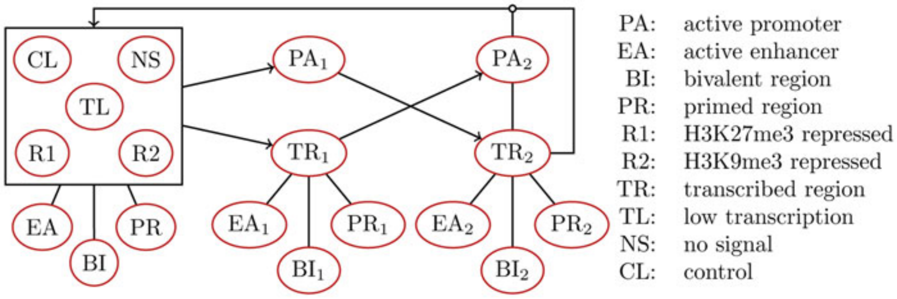

FIG. 2. ModHMM state diagram. Some chromatin states are represented by multiple hidden states. For instance, active enhancers (EA) are represented by hidden states EA, EA$_1$, and EA$_2$. If two states are connected by an undirected edge, transitions in both directions are allowed. Self-transitions are in general allowed but omitted in the figure. A box is used to group states that are fully connected. If an arrow points to the box, transitions to any of the states in that box are admissible. Crossing edges are connected if marked with a circle.

state $s \in \mathcal{S} = \{\text{PA}, \text{EA}, \text{BI}, \ldots\}$. The output of the classifiers are the genome-wide predictions of chromatin states, that is, prediction $c_t(s)$ yields the assessment of the expert for chromatin state s that genomic position t is in this state. In the supra-Bayesian approach, expert predictions are treated as observations and a separate model, the decision maker, is constructed to reach a final decision. Here, we decided to implement the decision maker as an HMM in which each chromatin state s is associated with one or more hidden states $s' \in \mathcal{S}' = \{\text{PA}, \text{EA}, \text{EA}_1, \text{EA}_2, \ldots\}$. For the emission model several choices would be conceivable. For instance, the family of beta distributions is frequently used to model probabilities. However, to reduce the number of parameters that must be estimated from data we decided to use a likelihood model that contains no free parameters and incorporates the classifier predictions as they are. We define the emission distribution of state s' in terms of the density function

$$f_{s'}(x) \propto x,$$

where $x = c_t(s)$ and s' is associated with chromatin state s.

2.1.3. *Transition rates.*

In typical HMMs, transition rates are estimated from data and reflect context-dependent state prevalences. The situation is different in our case, where classifiers are used to optimally discriminate between chromatin states and to account for prevalences. By naively integrating classifiers into an HMM, results of classification may get overruled by transition rates, making it more difficult to combine classifiers into a well-functioning model. For instance, whether a region is classified as active enhancer or primed should be mostly decided based on what classifier shows the higher probability (Fig. 3). Still, transition rates are valuable parameters and we use them to model the expected length of chromatin states.

Two types of constraints are imposed on the transition matrix Σ. First, the structure of the HMM forbids certain transitions resulting in entries that must remain zero during learning. Second, transition rates should only account for the average length of chromatin states. To accomplish this, it is necessary to constrain all nonzero off-diagonal entries within a row to share the same value. More specifically, ModHMM uses the transition matrix $\Sigma = (\sigma_{ij})$ with

$$\sigma_{ij} = \begin{cases} \delta_i & \text{if } i = j, \\ \nu_i & \text{if } i \neq j \text{ and transition is feasible}, \\ 0 & \text{otherwise}. \end{cases}$$

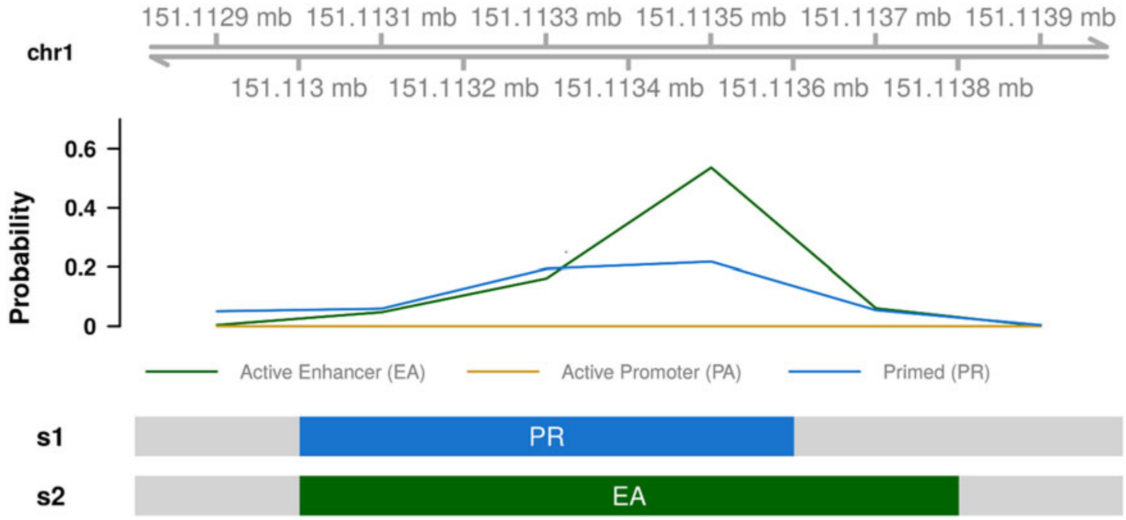

FIG. 3. Transition rates interfering with classifications. The figure shows a region with histone modification signals (omitted) of an active enhancer. The active enhancer classifier (EA) also shows the highest probability. The primed state classifier (PR) shows a smaller peak while the probability for an active promoter (PA) is very low. The first segmentation (s1) is computed with an unconstrained transition matrix. The region is classified as primed because transition rates overrule classifier assessments. The second segmentation (s2) has constrained transition rates and correctly classifies the region as active enhancer.

The parameter δ_i models the expected length of the ith chromatin state, whereas ν_i represents the transition rate into another state. In addition to the above constraints, we also assume that the expected length of chromatin states is context independent, that is, $\delta_i = \delta_j$ for $i, j \in \{\mathrm{EA}, \mathrm{EA}_1, \mathrm{EA}_2\}$, $i, j \in \{\mathrm{BI}, \mathrm{BI}_1, \mathrm{BI}_2\}$, $i, j \in \{\mathrm{PR}, \mathrm{PR}_1, \mathrm{PR}_2\}$, and $i, j \in \{\mathrm{TR}_1, \mathrm{TR}_2\}$. For each row i, diagonal entries δ_i and off-diagonal entries ν_i must be chosen such that the row sum is equal to 1. The constraints that are imposed on the transition matrix $\Sigma = (\sigma_{ij})$ complicate the estimation step, which led us to develop a modified Baum–Welch algorithm.

2.2. *Algorithm for estimating transition rates*

ModHMM incorporates predictions from a set of classifiers through a simple parameter-free likelihood model. Hence, the only parameters that must be estimated are the transition rates of the unobserved Markov process. Given the amount of data available for computing genome segmentations, we refrain from using highly sophisticated Bayesian estimation methods. Instead, we utilize the common maximum likelihood approach. As an instance of the expectation-maximization (EM) algorithm the Baum–Welch algorithm maximizes the likelihood function iteratively until a stationary point is found. To satisfy our constraints on the transition matrix $\Sigma = (\sigma_{ij})$ we employ a modified Baum–Welch algorithm that allows arbitrary equality constraints. Given a sequence of T observations, the updated transition matrix at iteration $n+1$ is given by

$$\Sigma_{n+1} = \arg\max_{\Sigma} \sum_{i,j} \xi_{ij} \log \sigma_{ij}, \quad \text{where} \quad \xi_{ij} = \sum_{t}^{T-1} \xi_{ij}(t),$$

and $\xi_{ij}(t)$ is the posterior probability of a transition from state i to j at position t computed using the transition matrix Σ_n. The optimization problem is solved by finding the stationary points of the Lagrangian

$$\mathcal{L}(\Sigma, \lambda) = \sum_{i,j} \xi_{ij} \log \sigma_{ij} - \sum_{i} \lambda_i \left[\sum_{j} \sigma_{ij} - 1 \right]$$

with multipliers λ_i. We explain the general solution on a small example with transition matrix

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & 0 & \sigma_q & \dots \\ \sigma_{21} & \sigma_{22} & \sigma_q & \sigma_q & \\ \vdots & & \ddots & & \vdots \end{bmatrix}$$

that contains two types of constraints. First, σ_{13} is set to zero so that transitions from state one to state three are not allowed. Second, all other transitions from the first two states to states three and four are constrained to have the same rate σ_q, that is, $\sigma_{14} = \sigma_{23} = \sigma_{24} = \sigma_q$. Finding the stationary points of $\mathcal{L}$ we obtain

$$\sigma_{ij} = \begin{cases} \frac{\xi_{ij}}{\lambda_i} & \text{if } (i, j) \in p = \{1, 2\} \times \{1, 2\} \\ \frac{\xi_q}{\sum_k |\sigma_q|_k \lambda_k} & \text{if } (i, j) \in q = \{(1, 4), (2, 3), (2, 4)\} \end{cases},$$

where $\xi_q = \sum_{(i,j) \in q} \xi_{ij}$. $|\sigma_q|_k$ denotes the number of σ_q entries in Σ at row k. In this example we have $|\sigma_q|_1 = 1$ and $|\sigma_q|_2 = 2$. The Lagrange multipliers are obtained as a solution to the constraints

$$\begin{aligned} \frac{\xi_{11}}{\lambda_1} + \frac{\xi_{12}}{\lambda_1} + \frac{|\sigma_q|_1 \xi_q}{\sum_k |\sigma_q|_k \lambda_k} + \ldots &= 1 \\ \frac{\xi_{21}}{\lambda_2} + \frac{\xi_{22}}{\lambda_2} + \frac{|\sigma_q|_2 \xi_q}{\sum_k |\sigma_q|_k \lambda_k} + \ldots &= 1, \end{aligned}$$

where dots mark additional constraints stemming from the remaining parts of the matrix Σ. Newton's root finding method (Gill et al., 1981) is used to solve this set of equations numerically.

2.3. *Default chromatin state classifiers*

ModHMM takes as input the genome-wide predictions of a set of classifiers. In principle, any type of classifier can be used; however, ModHMM implements a default classifier set to simplify usage. These engineered classifiers require no training data and consist of two layers. First, a single-feature classifier is constructed for each feature that determines the probability of enrichment at each genomic position. These single-feature classifiers are then combined into a set of naive Bayesian multifeature classifiers (Maron, 1961; Duda and Hart, 1973; Mitchell, 1997; Russell and Norvig, 2016), each specialized in the detection of a single chromatin state.

2.3.1. Single-feature classifiers. The purpose of a single-feature classifier is to assess the enrichment of a feature genome-wide. Since it assigns a probability to each genomic position, we may also interpret this step as a normalization step that decouples the definition of the engineered multifeature classifiers from the actual observations. Such classifiers are the basic ingredient of many peak calling methods, implementing a large variety of different statistical models (Wilbanks and Facciotti, 2010), ranging from Poisson (Mortazavi et al., 2008; Valouev et al., 2008) or local Poisson (Zhang et al., 2008) to HMMs (Spyrou et al., 2009).

Our approach differs in that we do not assume the same model for all features but rather account for the high heterogeneity. More specifically, we first compute the coverage along the genome in 200 bp bins (Section 3.5). The coverage values are then modeled by a feature-specific mixture distribution. Consider an event $\{X_t^{\varphi}=x\}$ with coverage value x from a feature $\varphi \in \{\text{ATAC}, \text{H3K4me1}, \ldots\}$ at bin t. We assume that

$$X_t^{\varphi} \sim \sum_{k \in F \cup B} \pi_k p_k,$$

where p_k is the kth component of the feature-specific mixture distribution with weight π_k. F and B partition the set $\{p_k\}$ into foreground $\{p_k | k \in F\}$ and background components $\{p_k | k \in B\}$. As an example, we take the ATAC-seq feature, which measures the accessibility of chromatin. The foreground component consists of two geometric distributions, whereas the background is modeled by a delta distribution positioned at zero, a Poisson and a geometric distribution (Fig. 4). Whether a component belongs to the foreground or background is a subjective choice and must be determined by visual inspection of the data. The parameters of the mixture distribution are estimated by maximum likelihood using the EM algorithm (Dempster et al., 1977).

Once a mixture distribution for a feature φ is determined, the probability that a given bin t of the genome with coverage value x is enriched is given by the posterior probability

$$q_t(\varphi) = \frac{\sum_{k \in F} \pi_k p_k(x)}{\sum_{k \in F \cup B} \pi_k p_k(x)} .$$

In this way, a single-feature classifier is constructed for all features. Every such classifier consists of a mixture of Poisson, geometric, and delta distributions.

2.3.2. Multifeature classifiers. Multifeature classifiers are defined as simple combinations of single-feature classifiers. Consider the case of active promoters that are known to be accessible and marked by H3K27ac and H3K4me3 as well as a high H3K4me3-to-me1 ratio (H3K4me3/1). A classifier for active promoters should assign high probabilities to regions where those three features co-occur. Therefore, a naive Bayesian promoter model is given by

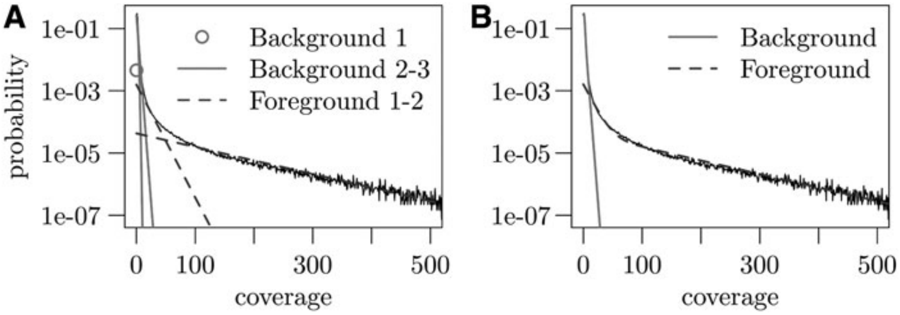

FIG. 4. Mixture distribution used to model ATAC-seq data. Both plots show the empirical distribution in black. **(A)** Components of the mixture distribution. **(B)** Foreground and background components are merged.

$$c_t(\mathrm{PA}) = q_t(\mathrm{ATAC}) \cdot q_t(\mathrm{H3K27ac}) \cdot q_t(\mathrm{H3K4me3}) \cdot q_t(\mathrm{H3K4me3/1}) \cdot \bar{q}_t(\mathrm{Control}),$$

where

$$\bar{q}_t(\mathrm{Control}) = 1 - q_t(\mathrm{Control})$$

enforces that no peak is observed in the control data set. The classifier can be improved by also considering the spatial structure of features. For instance, histone modifications are typically more broadly distributed than ATAC-seq peaks. For such features it is necessary to also consider surrounding bins and ask for the probability that any one of the bins is enriched. For a feature φ the probability that any one out of three adjacent bins is enriched is given by

$$q_{t-1:t+1} = q_{t-1} + \bar{q}_{t-1} \cdot q_t + \bar{q}_{t-1} \cdot \bar{q}_t \cdot q_{t+1},$$

where function arguments have been omitted for better readability. Some features may require a more detailed modeling of the spatial structure. For instance, H3K4me1 is symmetrically distributed around regulatory elements as opposed to H3K4me3 that shows a higher enrichment at promoters toward transcribed regions due to its role in preinitiation complex formation (Lauberth et al., 2013). The symmetric structure of a feature is captured by

$$s_{t-1:t+1} = q_{t-1} \cdot q_{t+1} + \overline{q_{t-1} \cdot q_{t+1}} \cdot q_t,$$

where

$$\overline{q_{t-1} \cdot q_{t+1}} = q_{t-1} \cdot \bar{q}_{t+1} + \bar{q}_{t-1} \cdot q_{t+1} + \bar{q}_{t-1} \cdot \bar{q}_{t+1}.$$

The classifier requires an enrichment at bins $t-1$ and $t+1$ or an enrichment at the center bin t.

In this manner, a multifeature classifier is constructed for every chromatin state. The classifiers are then assigned to states of the HMM, where some classifiers may also be shared among several states. This is, for instance, the case for the active enhancer states EA, EA$_1$, and EA$_2$, as well as the bivalent states BI, BI$_1$, and BI$_2$. A full specification of the classifiers is given in Table 1.

2.4. Supervised shape classifier

As an alternative to the engineered default classifiers we consider a supervised method that models the shape of features around chromatin states (Section 3.4). The method is inspired by DFilter (Kumar et al., 2013), a peak calling method that implements a Hotelling filter. In its essence, the method models foreground and background as multivariate normal distributions, sharing the same covariance matrix. A likelihood-ratio test is then used to call peaks (linear discriminant analysis). We modified the method and used several components for the background distribution. More specifically, at each genomic bin t we consider for all features the observations inside a window of size $2n+1$. Under the foreground model each random variable from the set $\{X^{\varphi}_{t+j}\}$ with $j \in \{-n, -n+1, \ldots, n\}$ is distributed as

Table 1. Multifeature Classifier Definitions

	PA	EA	PR		BI		TL	TR	CL	NS	R1	R2
ATAC	$\checkmark_c$	$\checkmark_c$	$\checkmark_c$	$\checkmark_c$			$\times_c$	$\times_c$		$\times_c$		
H3K27ac	$\checkmark_s$	$\checkmark_s$	$\times_a$	$\times_a$						$\times_c$		
H3K4me1		$\checkmark_{s,3}$		$\checkmark_{s,3}$		$\checkmark_{s,3}$				$\times_c$	$\times_c$	$\times_c$
H3K4me3	$\checkmark_a$		$\checkmark_a$		$\checkmark_a$		$\times_c$	$\times_c$		$\times_c$	$\times_c$	$\times_c$
H3K4me3/1	$\checkmark_a$	$\times_a$										
H3K27me3			$\times_a$	$\times_a$	$\checkmark_s$	$\checkmark_s$				$\times_c$	$\checkmark_c$	
H3K9me3										$\times_c$		$\checkmark_c$
RNA								$\checkmark_c$		$\times_c$		
RNA (low)							$\checkmark_c$					
Control	$\times_a$	$\times_a$	$\times_a$	$\times_a$	$\times_a$	$\times_a$			$\checkmark_c$	$\times_c$	$\times_c$	$\times_c$

$\checkmark_c$: bin t is enriched $[q_t]$; $\checkmark_a$: at least one bin out of $\{t-2, \ldots, t+2\}$ is enriched $[q_{t-2:t+2}]$; $\checkmark_s$: symmetric enrichments at bins $\{t-2, \ldots, t+2\}$ $[s_{t-2:t+2}]$; $\checkmark_{s,3}$: symmetric enrichments at bins $\{t-3, \ldots, t+3\}$ $[s_{t-3:t+3}]$; $\times_a$: no enrichment in all bins $i \in \{t-2, \ldots, t+2\}$ $[\bar{q}_{t-2:t+2}]$; $\times_c$: bin t shows no enrichment $[\bar{q}_t]$.

$$X_{t+j}^{\varphi} + 1 \sim \mathrm{LogNormal}(\mu_j^{\varphi}, \sigma_j^{\varphi}),$$

where the parameters $\{\mu_{jk}^{\varphi}\}$ and $\{\sigma_{jk}^{\varphi}\}$ determine the average shape and dispersion of features around the chromatin state. The background model is similar, but we consider a mixture of K components to model different shapes observed in the background data, that is,

$$Z_t \sim \mathrm{Categorical}(\pi)$$
$$X_{t+j}^{\varphi} + 1 | Z_t = k \sim \mathrm{LogNormal}(\mu_{jk}^{\varphi}, \sigma_{jk}^{\varphi}),$$

where the mixture weights π have dimension K. The parameters of the foreground and background model are estimated independently by maximum likelihood. Since the background model is a mixture distribution, we use the EM algorithm to maximize the likelihood function. The shape classifier computes the posterior probability that a given observation belongs to the foreground model.

3. RESULTS

We compared our method, equipped with its default set of engineered chromatin state classifiers, with two other segmentation methods, namely ChromHMM (Ernst and Kellis, 2012) and EpiCSeg (Mammana and Chung, 2015). ChromHMM is the most popular segmentation method, although it uses Bernoulli emission distributions for which the data must first be binarized into enriched and nonenriched regions. To also incorporate how strongly genomic regions are enriched, EpiCSeg uses negative multinomial emissions. Compared with the multinomial distribution, the negative multinomial better models the variability observed in most ChIP-seq data. We decided to omit a comparison with Segway, since its segmentations are even more fine-grained than those of ChromHMM and EpiCSeg (Hoffman et al., 2012a). We also omit a comparison with ChroModule (Won et al., 2013) since no software package was published by the authors. For all three methods a bin size of 200 bp was used. The ModHMM segmentation is computed as the most likely sequence of hidden states, that is, the Viterbi path. ChromHMM and EpiCSeg use the posterior decoding algorithm to compute segmentations.

We evaluated all three methods on ENCODE (ENCODE Project Consortium, 2012) data from mouse embryonic liver at day 12.5 and embryonic lung at day 16.5. For a first qualitative comparison, Figure 5 shows a small region within chromosome 1 that contains three actively transcribed genes, as indicated by the data. The ModHMM segmentation correctly detects the promoters and transcribed regions. In addition, there is a primed region located in one of the gene bodies. For ChromHMM and EpiCSeg the number of states was determined by maximizing classification performance (Section 3.1). The segmentations of both methods are highly fragmented and much more difficult to interpret. States that appear close to the promoter are also found at the primed region. This is due to the lack of an appropriate model for the spatial distribution of features around regulatory elements. Figure 6 shows state equivalences between the three methods. In general, there is a low overlap between states. Since we used more states for ChromHMM and EpiCSeg, it is clear that a single ModHMM state must be represented by multiple states from ChromHMM and EpiCSeg. However, there are also multiple recombinations, for instance, ChromHMM state 9 corresponds to ModHMM active promoter (PA) and enhancer (EA) states.

3.1. *Enhancer predictions*

To compare the predictive performance of segmentation methods we constructed a test set with active enhancers identified by the FANTOM consortium (Andersson et al., 2014). Enhancers are experimentally detected as origins of bidirectional capped transcripts using CAGE (Shiraki et al., 2003). We took all enhancers that showed at least 5 CAGE reads in liver at embryonic day 12 resulting in only 537 regions. With such a low detection threshold on the number of counts many false positives ought to be expected. Indeed, about half of the regions did not show the desired histone modification patterns. To filter false positives, we clustered the enhancer regions using deepTools2 (Ramírez et al., 2016). We dropped all clusters that either showed no histone marks or high levels of H3K4me3, resulting in 265 positive regions. The detection of active enhancers based on chromatin marks is difficult, mostly because promoters show a very similar pattern. Therefore, we constructed a test set consisting of 2650 regions, 1/10th are the filtered FANTOM enhancers, 8/10th promoters, and 1/10th random genomic regions.

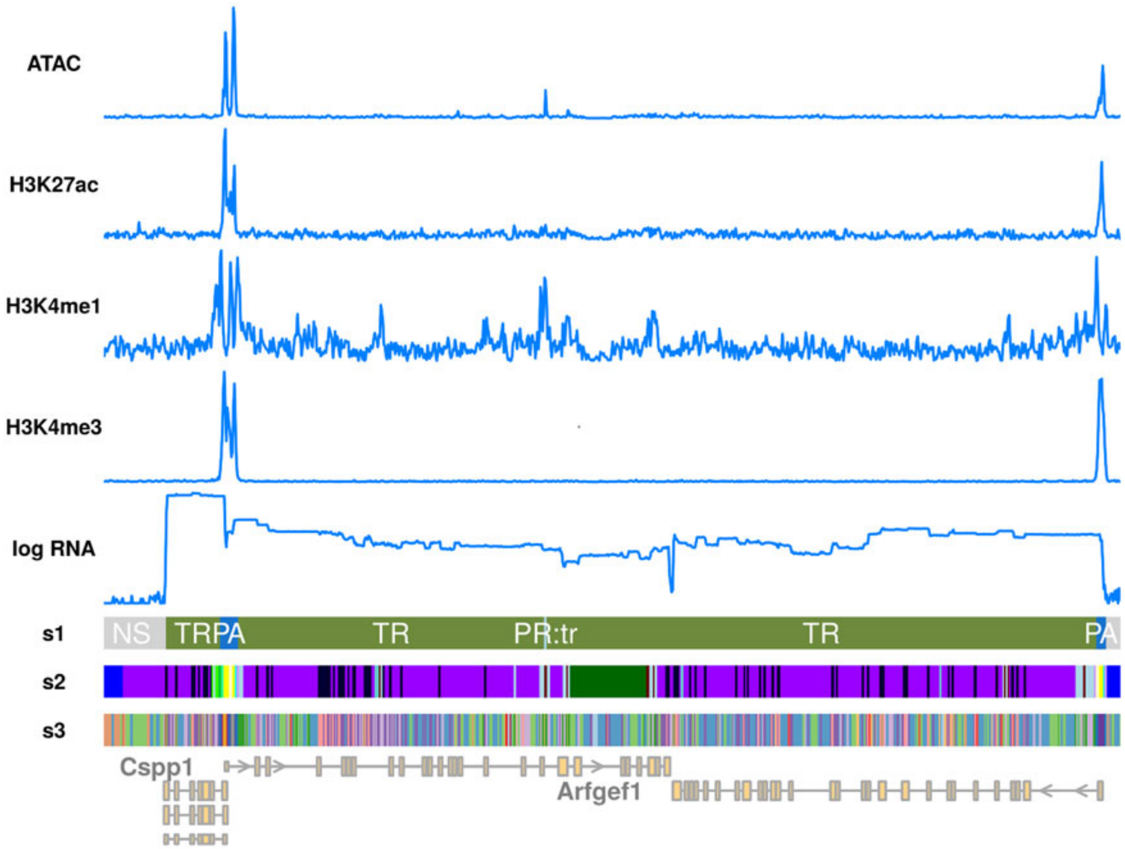

FIG. 5. Qualitative comparison of segmentation methods. s1: ModHMM, s2: ChromHMM, s3: EpiCSeg. The primed states of ModHMM within transcribed regions (PR_1, PR_2) are both abbreviated as PR:tr.

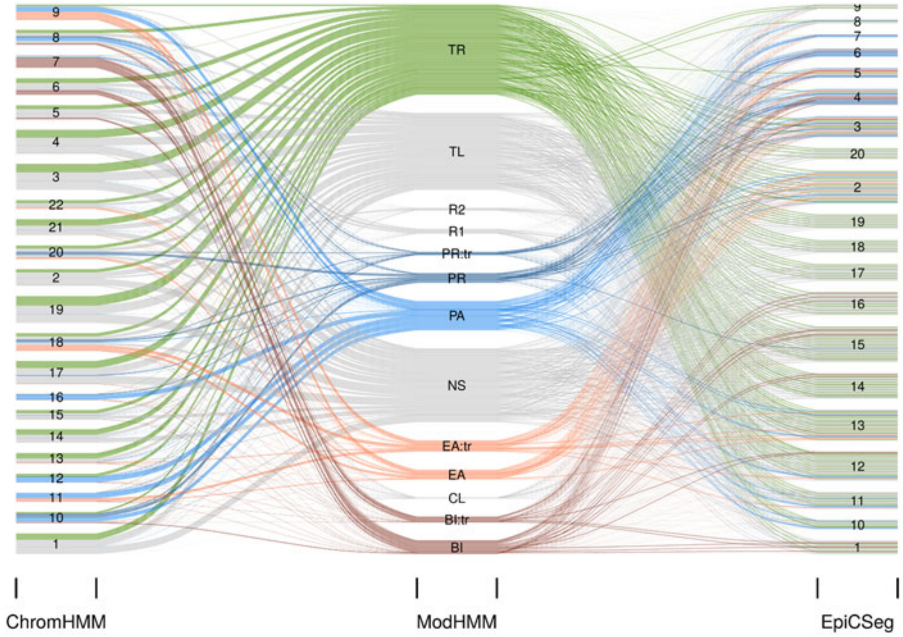

FIG. 6. State equivalences between ModHMM, ChromHMM, and EpiCSeg. Nodes represent the states of HMMs labeled either by state number (ChromHMM and EpiCSeg) or state name (ModHMM). Arrow widths indicate the fraction of genomic bins shared by two states. Active enhancer (EA_1, EA_2), primed (PR_1, PR_2), and bivalent (BI_1, BI_2) states within transcribed regions are, respectively, abbreviated as EA:tr, PR:tr, and BI:tr.

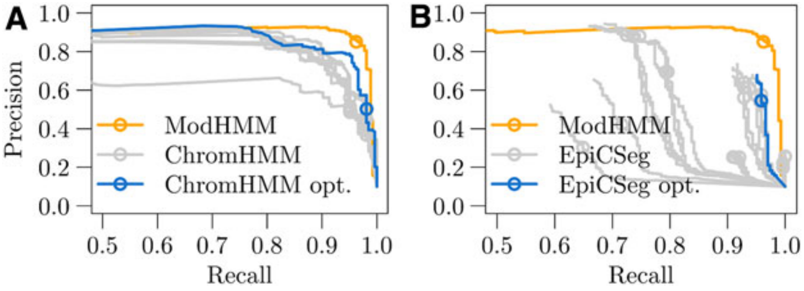

FIG. 7. Classification performances of active enhancer regions in mouse embryonic liver at day 12.5. **(A)** Performance of ModHMM and ChromHMM. Lines show performances evaluated using posterior marginals, whereas dots mark the performances of segmentations. For ChromHMM models with an even number of states between 10 and 30 were tested. The precision-recall curve is evaluated for several states and all possible combinations. For each model only the best curve is shown in gray. The optimal (opt.) curve is highlighted in blue. **(B)** Performance summary of ModHMM and EpiCSeg similar to **(A)**.

ModHMM has a well-defined enhancer state, whereas ChromHMM and EpiCSeg are unsupervised methods where states must be assigned a function after training. This assignment is often difficult especially for EpiCSeg where also the intensity of enrichment is modeled. To avoid wrong assignments, we consider for each model two to three putative enhancer states that are most abundant at the positive enhancer regions. Performance is then evaluated based on these states, including all possible combinations. The best performance is then reported, potentially giving ChromHMM and EpiCSeg a strong advantage over ModHMM.

Results are summarized in Figure 7. For all three methods we used posterior marginals of one or several states to compute precision-recall curves. In addition, we computed the classification performances of Viterbi paths. ModHMM shows the highest area under the precision-recall curve. The best ChromHMM model consists of 22 states and surprisingly outperforms the best EpiCSeg model with 20 states. The Viterbi path of ModHMM optimally balances precision and recall yielding the highest F-score. In contrast, especially the segmentations of ChromHMM show a poor balance of precision and recall with a maximum precision of $\sim 60\%$.

3.2. *Promoter state frequencies*

To understand why ModHMM performs better than ChromHMM and EpiCSeg, we looked at segmentations around active promoters. We used UCSC refGenes to obtain an initial set of transcription start sites (TSSs). Promoters are defined as 2 kbp windows around the TSS. From this set we took regions that have a clear ATAC-seq, H3K27ac, and RNA-seq signal. Regions enriched with H3K27me3 were filtered out. For each segmentation method, we computed at every position relative to the TSS the frequency of every state.

For ModHMM we observe a clear enrichment of the active promoter state (PA) around transcription start sites (Fig. 8). The active transcription state (TR) is flanking this region in most cases, whereas other states

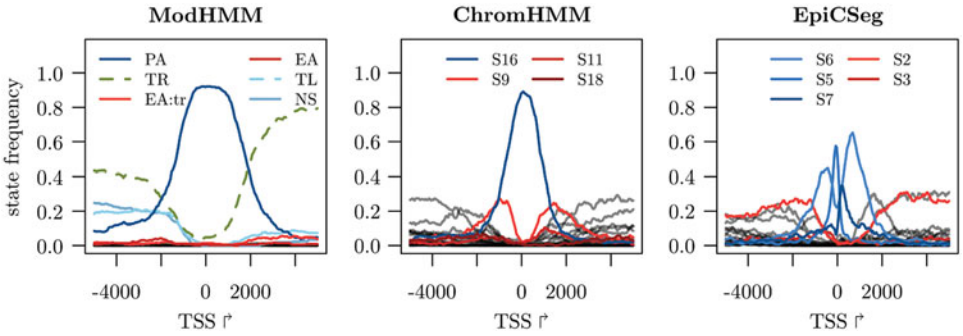

FIG. 8. State frequencies at promoters. Promoters that belong to genes on the reverse strand are inverted so that the gene body is right of the TSS. For ChromHMM and EpiCSeg, states that are frequently observed at enhancers are in red. The ModHMM state EA:tr refers to both active enhancer states (EA_1 and EA_2) inside transcribed regions.

are rarely observed. For ChromHMM active promoters are modeled by state 16, which represents enrichment in ATAC-seq, H3K27ac, H3K4me3, and RNA-seq. However, the region represented by this state is much narrower and it is frequently flanked by a diverse set of states. One of them is state 9, which models enrichment in ATAC-seq, H3K27ac, H3K4me1, and H3K4me3. It is also frequently found at enhancers that show enrichment in H3K4me3 above the binarization threshold set by ChromHMM. The situation is similar for EpiCSeg; however, the promoter is fragmented into several states modeling different levels of ATAC-seq and H3K4me3 enrichment. State 2 is frequently flanking promoters, which also occurs at enhancers since it models enrichment in ATAC-seq, H3K27ac, and H3K4me1, but low enrichment in H3K4me3. Peaks of H3K4me3 tend to be more localized than H3K4me1 peaks, so that regions close to promoters typically show characteristics of enhancers. Both ChromHMM and EpiCSeg do not model the spatial distribution of features around promoters and enhancers and, therefore, often fail to correctly discriminate between them.

3.3. Bivalent state

During the development of ModHMM, we observed that the H3K4me1-to-me3 ratio has low predictive power for discriminating between bivalent promoters and poised enhancers. This led us to represent both chromatin states by a single bivalent state. To quantify this observation, we consider all bivalent regions in the ModHMM segmentation of mouse embryonic liver at day 12.5. The H3K4me1-to-me3 ratio is then used to separate promoters from the remaining regions (i.e., putative poised enhancers). All bivalent regions overlapping annotated UCSC refGene promoters (500 bp regions around transcription start sites) are defined as true positives. This leads to an area under the precision recall curve (PR-AUC) of ∼0.84. The minimal PR-AUC achieved by a random classifier is ∼0.63. As a comparison, we took all regions of the ModHMM segmentation that are either labeled as active promoter or enhancer. Here, the same procedure leads to a PR-AUC of ∼0.96, whereas a random classifier achieves a performance of 0.46.

3.4. Supervised cross-tissue enhancer predictions

The modular structure of ModHMM allows us to gradually improve the segmentation by replacing single classifiers. Here, we graft a supervised classifier into ModHMM that models the shape of features at enhancers (Section 2.4). We use the FANTOM enhancers from embryonic liver at day 12.5 as training set and test the method on lung at day 16.5, for which a validation set has been generated in the same way (Section 3.1).

Typically there is much variability between data sets, which prevent a direct application of the model to different tissues. In most cases it would be necessary to re-estimate the parameters of the single-feature classifiers. Owing to the uniform experimental protocols and processing steps of ENCODE data we may, however, skip this step. Still, there remain some statistical differences between the data sets that we remove by quantile normalizing (Amaratunga and Cabrera, 2001; Bolstad et al., 2003) the coverage values of each feature.

To establish a baseline, we first tested how well ModHMM with its default chromatin state classifiers performs across tissues. We also trained a ChromHMM and EpiCSeg model on data from embryonic lung at day 16.5. However, we kept for both methods the number of states that was optimal for the liver data set. Compared with the liver data, the classification performance of ModHMM dropped slightly, but also the optimal combination of states for ChromHMM and EpiCSeg showed a much lower classification performance (Fig. 9A). Next, we tested the classification performance of the supervised shape classifier for which we used a window size of 10 bins and ATAC-seq, H3K27ac, H3K4me1, and H3K4me3/1 as well as the control data as features. In general, the performance of the shape classifier is better than the engineered classifier (Fig. 9B). Integrated into ModHMM the performance stays approximately the same, showing that other classifiers do not interfere.

3.5. Data preparation

ENCODE BAM accession numbers are given in Table 2. For embryonic lung at day 16.5 no ATAC-seq BAM files were available. We downloaded fastq files from ENCODE with accession nos. ENCFF577XAL, ENCFF224TAO, ENCFF872IZI, and ENCFF427HTC. Bowtie2 (Langmead and Salzberg, 2012) was used to align reads to the reference genome mm10. Picard and samtools (Li et al., 2009) were used to mark and filter duplicates.

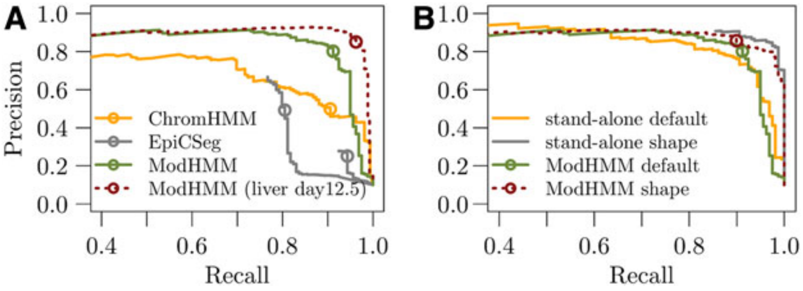

FIG. 9. Cross-tissue predictive performance. (**A**) Predictive performance of ModHMM, ChromHMM, and EpiCSeg on data from embryonic lung at day 16.5. (**B**) ModHMM performance on lung at day 16.5 with supervised shape classifier.

ENCODE data are available as either single- or paired-end. ATAC-seq data were generated using paired-end sequencing; however, since we are only interested in accessible regions we treated it as single end. For histone modifications single-end sequencing was used. We processed this data by first estimating the fragment length with a cross-correlation analysis of forward and reverse strand reads. Afterward, reads are extended in 3′ direction to match the estimated fragment length. No special processing of RNA-seq data was required.

The coverage C_n of each feature is computed by first tiling the genome into a set of bins of size n. Given a set of reads R, for each $r \in R$ the value of all bins that overlap with r is incremented by the number of overlapping nucleotides. The final coverage is the total number of overlapping nucleotides for all reads per bin size n. The advantage of this binning scheme is that a coverage C_m with bin size m can be computed by increasing the bin size of another coverage C_n as long as m is a multiple of n.

The feature H3K4me3/1 was computed from the coverages of H3K4me1 and H3K4me3. At bin t the ratio is given by

$$z_t = \text{round}\left(10 \cdot \frac{x_{t-1} + x_t + x_{t+1}}{y_{t-1} + y_t + y_{t+1} + 1} \frac{\sum_s y_s}{\sum_s x_s}\right),$$

where $\{x_t\}$ are the coverage values of H3K4me3 and $\{y_t\}$ those of H3K4me1.

4. DISCUSSION

Traditional genome segmentation methods, such as ChromHMM, EpiCSeg, or Segway, are unsupervised methods and can be used to detect known and unknown patterns in genomics data. They have been extensively used in the past to analyze the epigenetic landscape of a large variety of cell types and tissues (Hoffman et al., 2012b; Kundaje et al., 2015; Ernst and Kellis, 2017). However, nowadays much is known about the epigenetic landscape and the features that mark regulatory elements. This extensive knowledge questions, at least to some extent, the traditional approach to genome segmentation. Instead, we used this knowledge to construct a segmentation method that outperforms the traditional methods in several aspects.

TABLE 2. ENCODE BAM FILE ACCESSION NUMBERS

	Embryonic liver day 12.5	*Embryonic lung day 16.5*
ATAC	ENCFF929LOH, ENCFF848NLJ	—
H3K27ac	ENCFF524ZFV, ENCFF322QGS	ENCFF431CSN, ENCFF149TQU
H3K27me3	ENCFF811DWT, ENCFF171KAM	ENCFF079QHR, ENCFF788UXL
H3K4me1	ENCFF788JMC, ENCFF340ACH	ENCFF151AXQ, ENCFF440VFN
H3K4me3	ENCFF211WGC, ENCFF587PZE	ENCFF844DSD, ENCFF437ESL
H3K9me3	ENCFF293UCG, ENCFF777XFH	ENCFF740KNN, ENCFF380FIJ
RNA	ENCFF405LEY, ENCFF627PCS	ENCFF502KFQ, ENCFF371ONT
Control	ENCFF865QGZ, ENCFF438RYK	ENCFF018HYW, ENCFF855ISH

ModHMM has a higher prediction accuracy and the segmentations show a better balance of precision and recall. With each hidden state of ModHMM a classifier is associated that detects a well-defined chromatin state. This leads to segmentations that are superior in qualitative aspects. Functional elements, such as active promoters or enhancers, are typically contained in a single segment, which is not the case for ChromHMM and EpiCSeg.

Inspired by the supra-Bayesian approach, ModHMM integrates predictions of a set of experts or classifiers. Using the output of classifiers as input to the HMM has certain advantages over classical HMMs that model observations directly. For instance, a classifier may cherry-pick only a subset of the available data. The coverage of RNA-seq reads in a single genomic bin already provides enough information to decide whether the region is transcribed. In contrast, classification of active promoters and enhancers requires data from multiple features and several surrounding bins.

ModHMM uses a default set of engineered classifiers to detect chromatin states. The basis of which is a single-feature enrichment analysis with a mixture model tailored to each feature. This is unique to ModHMM as most peak calling methods implement a single model. Applying ModHMM to a new data set requires to perform the enrichment analysis de novo. Alternatively, ModHMM may quantile normalize a new data set to a known reference for which single-feature models already exist. We have shown on different tissues that this approach leads to good results. Unlike ChromHMM and EpiCSeg, ModHMM has well-defined hidden states that do not change when applied across different cell types or tissues. This makes ModHMM ideal for differential analysis.

Compared with traditional segmentation methods, ModHMM is much more flexible and provides many leverage points to construct high-quality segmentations. For instance, any of the chromatin state classifiers from the default set can be replaced by more accurate alternatives, allowing to incorporate predictions from state-of-the-art methods such as REPTILE (He et al., 2017). To improve a given segmentation, ModHMM allows visual inspection of all classifier predictions, which may serve as a powerful tool to decide which classifiers must be replaced. The software is freely available at (https://github.com/pbenner/modhmm).

ACKNOWLEDGMENTS

We thank Anna Ramisch, Tobias Zehnder, and Verena Heinrich for their comments on the article and many inspiring discussions.

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no competing financial interests.

FUNDING INFORMATION

P.B. was supported by the German Ministry of Education and Research (BMBF, grant no. 01IS18037G).

REFERENCES

Amaratunga, D., and Cabrera, J. 2001. Analysis of data from viral DNA microchips. *J Am Stat Assoc* 96, 1161–1170.

Andersson, R., Gebhard, C., Miguel-Escalada, I., et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* 507, 455.

Barski, A., Cuddapah, S., Cui, K., et al. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* 129, 823–837.

Bolstad, B.M., Irizarry, R.A., Åstrand, M., and Speed, T.P. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19, 185–193.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., et al. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nat Methods* 10, 1213.

Buenrostro, J.D., Wu, B., Chang, H.Y., and Greenleaf, W.J. 2015. Atac-seq: A method for assaying chromatin accessibility genome-wide. *Curr Protoc Mol Biol* 109, 21–29.

Burge, C., and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 268, 78–94.

Calo, E., and Wysocka, J. 2013. Modification of enhancer chromatin: What, how, and why? *Mol Cell* 49, 825–837.

Cappé, O., Moulines, E., and Rydén, T. 2005. *Inference in Hidden Markov Models*, volume 6. Springer, New York, NY, USA.

Coetzee, S.G., Ramjan, Z., Dinh, H.Q., et al. 2018. Statehub-statepaintr: Rapid and reproducible chromatin state evaluation for custom genome annotation. *F1000Research* 7.

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., et al. 2010. Histone h3k27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* 107, 21931–21936.

Dempster, A.P., Laird, N.M., and Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B (Methodol)* 1–38.

Duda, R.O., and Hart, P.E. 1973. *Pattern Classification and Scene Analysis*. A Wiley-Interscience Publication, New York.

Ernst, J., and Kellis, M. 2012. ChromHMM: Automating chromatin-state discovery and characterization. *Nat Methods* 9, 215.

Ernst, J., and Kellis, M. 2017. Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc* 12, 2478.

Galassi, U., Giordana, A., and Saitta, L. 2007. Structured hidden Markov model: A general framework for modeling complex sequences, 290–301. In *Congress of the Italian Association for Artificial Intelligence*. Springer, New York, NY, USA.

Gelfand, A.E., Mallick, B.K., and Dey, D.K. 1995. Modeling expert opinion arising as a partial probabilistic specification. *J Am Stat Assoc* 90, 598–604.

Genest, C., and Zidek, J.V. 1986. Combining probability distributions: A critique and an annotated bibliography. *Stat Sci* 1, 114–135.

Gill, P.E., Murray, W., and Wright, M.H. 1981. *Practical Optimization*. Academic Press, London, UK.

Gorkin, D., Barozzi, I., Zhang, Y., et al. 2017. Systematic mapping of chromatin state landscapes during mouse development. *bioRxiv* 166652.

He, Y., Gorkin, D.U., Dickel, D.E., et al. 2017. Improved regulatory element prediction based on tissue-specific local epigenomic signatures. *Proc Natl Acad Sci U S A* 114, E1633–E1640.

Heintzman, N.D., Stuart, R.K., Hon, G., et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* 39, 311.

Heinz, S., Romanoski, C.E., Benner, C., and Glass, C.K. 2015. The selection and function of cell type-specific enhancers. *Nat Rev Mol Cell Biol* 16, 144.

Hoffman, M.M., Buske, O.J., Wang, J., et al. 2012a. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* 9, 473.

Hoffman, M.M., Ernst, J., Wilder, S.P., et al. 2012b. Integrative annotation of chromatin elements from encode data. *Nucleic Acids Res* 41, 827–841.

Jacobs, R.A. 1995. Methods for combining experts' probability assessments. *Neural Comput* 7, 867–888.

Koch, F., Fenouil, R., Gut, M., et al. 2011. Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nat Struct Mol Biol* 18, 956.

Kumar, V., Muratani, M., Rayan, N.A., et al. 2013. Uniform, optimal signal processing of mapped deep-sequencing data. *Nat Biotechnol* 31, 615.

Kundaje, A., Meuleman, W., Ernst, J., et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317.

Kuzmichev, A., Nishioka, K., Erdjument-Bromage, H., et al. 2002. Histone methyltransferase activity associated with a human multiprotein complex containing the enhancer of zeste protein. *Genes Dev* 16, 2893–2905.

Langmead, B., and Salzberg, S.L. 2012. Fast gapped-read alignment with bowtie 2. *Nat Methods* 9, 357.

Lauberth, S.M., Nakayama, T., Wu, X., et al. 2013. H3k4me3 interactions with taf3 regulate preinitiation complex assembly and selective gene activation. *Cell* 152, 1021–1036.

Li, H., Handsaker, B., Wysoker, A., et al. 2009. The sequence alignment/map format and samtools. *Bioinformatics* 25, 2078–2079.

Lindley, D. 1982. The improvement of probability judgements. *J R Stat Soc (General)* 117–126.

Lindley, D. 1985. Reconciliation of discrete probability distributions, 375–390. *In* Bernardo, J., DeGroot, M., Lindley, D., and Smith, A., eds., *Bayesian Statistics 2: Proceedings of the Second Valencia International Meeting*. Valencia University Press, Valencia, Spain.

Lindley, D.V., Tversky, A., and Brown, R.V. 1979. On the reconciliation of probability assessments. *J R Stat Soc A (General)* 146–180.

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 Bio* 16, 151.

Margueron, R., and Reinberg, D. 2011. The polycomb complex prc2 and its mark in life. *Nature* 469, 343.

Maron, M.E. 1961. Automatic indexing: An experimental inquiry. *JACM* 8, 404–417.

Mitchell, T.M. 1997. *Machine Learning*. McGraw-Hill, Boston, MA.

Mortazavi, A., Williams, B.A., McCue, K., et al. 2008. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Methods* 5, 621.

Rabiner, L.R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE* 77, 257–286.

Ramírez, F., Ryan, D.P., Grüning, B., et al. 2016. deeptools2: A next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* 44, W160–W165.

Russell, S.J., and Norvig, P. 2016. *Artificial Intelligence: A Modern Approach*. Pearson Education Limited, Malaysia.

Saksouk, N., Simboeck, E., and Déjardin, J. 2015. Constitutive heterochromatin formation and transcription in mammals. *Epigenetics Chromatin* 8, 3.

Shiraki, T., Kondo, S., Katayama, S., et al. 2003. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci U S A* 100, 15776–15781.

Spyrou, C., Stark, R., Lynch, A.G., and Tavarà, S. 2009. Bayespeak: Bayesian analysis of ChIP-seq data. *BMC Bioinformatics* 10, 299.

Valouev, A., Johnson, D.S., Sundquist, A., et al. 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-seq data. *Nat Methods* 5, 829.

Wagner, E.J., and Carpenter, P.B. 2012. Understanding the language of lys36 methylation at histone h3. *Nat Rev Mol Cell Biol* 13, 115.

Wilbanks, E.G., and Facciotti, M.T. 2010. Evaluation of algorithm performance in chip-seq peak detection. *PLoS One* 5, e11471.

Won, K.-J., Zhang, X., Wang, T., et al. 2013. Comparative annotation of functional regions in the human genome using epigenomic data. *Nucleic Acids Res* 41, 4423–4432.

Zhang, Y., Liu, T., Meyer, C.A., et al. 2008. Model-based analysis of chip-seq (MACS). *Genome Biol* 9, R137.

Address correspondence to:
Dr. Philipp Benner
Department of Computational Molecular Biology
Max Planck Institute for Molecular Genetics
Ihnestraße 73
Berlin 14195
Germany

E-mail: benner@molgen.mpg.de