Published in final edited form as:

J Am Stat Assoc. 2014 January 1; 109(505): 78–94. doi:10.1080/01621459.2013.869222.

Some Statistical Strategies for DAE-seq Data Analysis: Variable Selection and Modeling Dependencies among Observations

Naim U. Rashid*, Wei Sun**, and Joseph G. Ibrahim***

Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, U.S.A

Abstract

In DAE (DNA After Enrichment)-seq experiments, genomic regions related with certain biological processes are enriched/isolated by an assay and are then sequenced on a high-throughput sequencing platform to determine their genomic positions. Statistical analysis of DAE-seq data aims to detect genomic regions with significant aggregations of isolated DNA fragments (“enriched regions”) versus all the other regions (“background”). However, many confounding factors may influence DAE-seq signals. In addition, the signals in adjacent genomic regions may exhibit strong correlations, which invalidate the independence assumption employed by many existing methods. To mitigate these issues, we develop a novel Autoregressive Hidden Markov Model (AR-HMM) to account for covariates effects and violations of the independence assumption. We demonstrate that our AR-HMM leads to improved performance in identifying enriched regions in both simulated and real datasets, especially in those in epigenetic datasets with broader regions of DAE-seq signal enrichment. We also introduce a variable selection procedure in the context of the HMM/AR-HMM where the observations are not independent and the mean value of each state-specific emission distribution is modeled by some covariates. We study the theoretical properties of this variable selection procedure and demonstrate its efficacy in simulated and real DAE-seq data. In summary, we develop several practical approaches for DAE-seq data analysis that are also applicable to more general problems in statistics.

Keywords

High-throughput Sequencing; Hidden Markov Model; Mixture Regression; Autoregressive modeling; Variable Selection

1. INTRODUCTION

High-throughput Sequencing (HTS) technologies are quickly gaining popularity in biomedical research. In this paper, we focus on the application of HTS to detect regions of the genome related with certain biological processes. In such applications, DNA pertaining to these regions are enriched by a biological assay, sequenced on a HTS platform, and then mapped to the reference genome to determine their genomic locations. Examples of these assays include Chromatin Immunoprecipitation (ChIP) that isolates regions containing protein-DNA interactions or epigenetic marks, DNase that selects for DNase I hypersensitive sites, and Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE) that isolates genomic regions of unpackaged chromatin. We collectively refer to these experiments as DAE(DNA After Enrichment)-seq experiments. Most DAE-seq experiments

* naim@unc.edu

** weisun@email.unc.edu

*** ibrahim@bios.unc.edu

aim to determine the locations of active regulatory elements, which provide important insight into gene regulation and molecular mechanisms underlying phenotypic outcomes, such as complex human diseases.

A typical DAE-seq study consists of four steps. First, the DNA sample is prepared and fragmented, where DNA fragments related with the biological activity of interest are enriched. Then, the first 30–100 base-pairs of one or both ends of each isolated DNA fragment are sequenced on a HTS platform, where the sequenced portions of each fragment are termed as “reads”. Next, the likely genomic position of each read is determined by short read alignment software in a process called “read mapping”. Finally, statistical and computational methods are employed to detect genomic regions with strong local aggregations of mapped DAE-seq reads, referred to as “enriched regions”. All other regions are referred to as “background”. Recent model-based methods for DAE-seq analysis [Kuan et al., 2011, Rashid et al., 2011, Spyrou et al., 2009, Ji et al., 2008, Qin et al., 2010] typically summarize local DAE-seq read density using “window read counts”, defined as the number of (single or paired-end) reads falling into consecutive windows of fixed size tiling the genome. The statistical goal of such methods is to identify regions (i.e., a group of consecutive windows) with window read counts significantly higher than the rest of the genome and/or a matching control. This summary is by nature a series of counts, and therefore the Poisson or Negative Binomial distribution is often used to model these counts and to dissect enriched and background regions.

There are several challenges in the analysis of DAE-seq data. First, several confounding factors may influence DAE-seq read density across the genome. Therefore, it is important to adjust for the effects of these factors, especially when a matching control dataset is not available [Rashid et al., 2011]. Examples of these factors include the local percentage of G and C nucleotides (“GC content”), the ability to accurately assign reads to a particular region of the genome (“mappability”), and the presence of local DNA copy number alterations [Rozowsky et al., 2009, Kuan et al., 2011, Rashid et al., 2011]. It may also be of interest to determine relationships between DAE-seq signal and multiple biological factors, such as the presence of protein-DNA interaction sites. However, when the number of factors (covariates) is large, it is challenging to choose the best subset of them to model DAE-seq data since the relevant set of covariates may be different for background and enriched regions.

Next, window read counts from DAE-seq data are often serially correlated. This correlation may simply be due to the dependence of underlying states of adjacent windows. For example, an enriched region may cover several consecutive windows in certain data types. However, we have noticed that given the underlying states, nearby windows’ read counts may still exhibit moderate to strong autocorrelation, even if they are from non-overlapping windows (Figure 1). This autocorrelation may be due to other covariates that are either unmeasurable or not included in the analysis, for example, DNA characteristics other than GC content or some bias due to the sequencing technique. Explicitly modeling this autocorrelation may explain a greater proportion of the variation in observed window read counts and may lead to more accurate estimates of the effects of other covariates as well as more accurate detection of enriched regions.

Several methods have been introduced to utilize Hidden Markov Models (HMMs) to account for the dependence between underlying states and identify enriched regions in DAE-seq data [Xu et al., 2008, Spyrou et al., 2009, Qin et al., 2010], where the transitions between latent states are explicitly modeled and the window read counts are assumed to be conditionally independent given the underlying states. One drawback of these approaches is that the confounding covariates, such as GC content and mappability, have not been

incorporated into the HMM. In addition, potential autocorrelation of adjacent windows given the underlying states is ignored. A few methods utilizing Finite Mixtures of Regression Models [Kuan et al., 2011, Rashid et al., 2011] have been proposed to incorporate the effects of multiple covariates to identify enriched regions. Unfortunately, these methods ignore any dependence between adjacent windows' read counts. Most notably, when the number of covariates is large, no computationally efficient method exists to automatically select state-specific covariates for HMMs where the observations are non-independent.

To address these challenges, we develop an Autoregressive Hidden Markov Model (AR-HMM) with covariates for DAE-seq data analysis. We derive a novel EM algorithm to estimate model parameters and we show that our method achieves better performance in the detection of enriched regions in simulated and real DAE-seq datasets. We also introduce a computationally efficient penalized maximum likelihood estimation procedure to perform state-specific variable selection, and establish the conditions for the existence, sparsity, and asymptotic normality of the penalized maximum likelihood estimates for a general class of penalty functions. We demonstrate the performance of this procedure in simulation studies, and apply it to discover a subset of 40 transcription factors whose protein-DNA interaction profiles are associated with a well-studied histone modification mark. In summary, we provide several practical solutions to challenges in DAE-seq analysis with broader applicability to other areas of statistics.

2. BACKGROUND

2.1 DAE-seq data analysis using Finite Mixtures of Regression Models

Consider a random sample of n responses $Y_1, \dots, Y_n$ from a Finite Mixture of Regressions Model (FMR) such that for each realization y_i

$$p(y_i | \mathbf{X}, \Psi_F) = \sum_{k=1}^K \pi_k f_k(y_i | X_{ik}, \beta_k, \varphi_k), \quad (1)$$

where K is the number of mixture components, $\mathbf{X}$ is an $n \times p$ matrix that includes the values of p covariates, $X_k \in \mathbb{R}_{n \times p_k}$ contains p_k columns of $\mathbf{X}$ that correspond to the p_k covariates pertaining to component k , $X_{ik} \in \mathbb{R}_{1 \times p_k}$ is the i^{th} row of X_k , $\Psi_F = (\beta^T, \phi^T, \pi^T)^T$,

$\beta = (\beta_1^T, \dots, \beta_K^T)^T$ where β_k is a $p_k \times 1$ vector of regression coefficients for component k , $\phi = (\phi_1, \dots, \phi_K)^T$ where ϕ_k is the dispersion parameters for the k -th component, and $\pi = (\pi_1, \dots,$

$\pi_K)^T$ is the set of prior probabilities of component membership such that $\sum_{k=1}^K \pi_k = 1$ and $\pi_k > 0$. Also, $f_k(y_i | X_{ik}, \beta_k, \phi_k)$ is the conditional density that y_i is generated from mixture component k with mean μ_{ik} and link function $h(\cdot)$ such that $h(\mu_{ik}) = X_{ik}\beta_k$. Denote the underlying mixture component for window i by Z_i where $Z_i = 1, \dots, K$. Under the assumptions of the FMR, we have $Z_i \perp Z_j$ and $y_i | Z_i \perp y_j | Z_j$ for $1 \leq i \neq j \leq n$. Given $\mathbf{X}$ and $\hat{\Psi}_F$, the posterior probability that window i belongs to component k can be computed and utilized for classification purposes [McLachlan, 1997]. In DAE-seq data analysis, each chromosome is typically modeled separately. Therefore the sample size of this problem is the number of windows spanning a chromosome, which may range from 100,000 to almost a million depending on the chosen window length (typically 50–500 bp) and chromosome size.

FMR-based methods such as Kuan et al. [2011] and Rashid et al. [2011] utilize $K = 2$ Negative Binomial mixture components pertaining to the background and enriched regions of DAE-seq data. In addition, Rashid et al. [2011] assumed an additional component to

account for potential zero-inflation in window read counts, whereas Kuan et al. [2011] modeled zero-inflation through a binary latent variable in the background component. These FMR-based approaches can flexibly account for the effects of multiple covariates that influence the window read counts in background and/or enriched regions. However, they ignore the dependence that may exist between adjacent windows, which may be due to dependence of underlying components or dependence of observations given underlying components. As a result, *ad-hoc* approaches were required to detect broader enriched regions for epigenetic marks [Rashid et al., 2011].

2.2 Variable Selection via Penalized Likelihood for FMR

In previous work involving FMRs and their applications to DAE-seq data analysis, Rashid et al. [2011] employed all-subset selection coupled with BIC [Schwarz, 1978] to select the best set of covariates for each mixture component. This approach is not computationally feasible when the number of covariates p is large, especially in the mixture distribution case where the number of possible models is 2^p [Khalili and Chen, 2007].

An enormous amount of statistical literature has been devoted to variable selection by penalized regression or penalized likelihood, and different types of penalty functions have been developed including the LASSO [Tibshirani, 1996], SCAD [Fan and Li, 2001], adaptive LASSO [Zou, 2006], MCP [Zhang, 2010], Log penalty [Friedman, 2008] among many others. Khalili and Chen [2007] have introduced variable selection via penalized likelihood in FMRs. They developed an EM algorithm to maximize the penalized FMR likelihood and showed that the Penalized Maximum Likelihood Estimate (PMLE) in the M-step of the EM algorithm can achieve the “oracle property”, where the zero coefficients are estimated to be zero with probability approaching to one and the non-zero coefficients are unbiasedly and efficiently estimated as if the “true” submodel is known [Fan and Li, 2001].

We extend the results of Khalili and Chen [2007] to establish an efficient variable selection procedure (EM + coordinate descent algorithm) in the context of Hidden Markov Models (HMMs) where the emission probability of each state is modeled by a set of covariates. We derive the asymptotic properties of the PMLE for the M-step of the algorithm and evaluate this algorithm using both simulations and real data analysis.

2.3 Accounting for Serial Dependence in Generalized Linear Models

Generalized linear models that account for serial correlation in observations fall into two categories: parameter-driven and observation-driven [Cox et al., 1981]. Parameter-driven models assume that the dependence between subsequent observations is controlled by a latent process that induces the correlation. For example, Zeger [1988] modeled a time series of counts, denoted by y_t , by a log-linear model conditioning on a latent process ε_t , such that $u_t = E(y_t | \varepsilon_t) = \exp(x_t' \beta) \varepsilon_t$ and $\text{var}(y_t | \varepsilon_t) = u_t$. The correlations among y_t 's are induced by the correlations among ε_t 's. In contrast, observation-driven models specify the conditional distribution of y_t as a function of past observations $y_{t-1}, \dots, y_1$. For example, an autoregressive (AR) model is an example of observation-driven model. Zeger and Qaqish [1988] introduced a Poisson generalized linear AR model, which, in the case of AR(1), has the following link function

$$\log(\mu_i) = X_i \beta + \nu \{ \log(y_{i-1} + c) - \log[\exp(X_{i-1} \beta) + c] \}, \quad (2)$$

where X_i is the i^{th} row of $\mathbf{X}$, i.e., the covariates' values for the i^{th} sample, β is a $p \times 1$ vector of regression coefficients, ν is the auto-correlation coefficient, and $0 < c < 1$ is used to avoid taking log of a zero.

Estimation for parameter-driven models is computationally difficult, especially in longer time series [Davis et al., 2003], making them less desirable choices in DAE-seq data analysis. Therefore, we utilize an observation-driven approach. Denote the data from the prior observation as $F_{i-1} = (X_{i-1}, y_{i-1})$. The model of Zeger and Qaqish [1988] assumes $\mu_i = E[Y_i|F_{i-1}]$ and $h(\mu_i) = X_i\beta + \text{vg}(F_{i-1})$, where $h()$ is a link function. We generalize the model of Zeger and Qaqish [1988] to an observation-driven autoregressive-HMM (AR-HMM) with K states. We assume an AR(1) dependence, which is reasonable for DAE-seq data. Let $Z_i = 1, \dots, K$ be a random variable of the underlying state of the i -th observation, and thus $Z = (Z_1, \dots, Z_n)$ are the random variables for the state path. Given a particular instance of state path, denoted by $z = (z_1, \dots, z_n)$, we have $g(F_{i-1}, z) = \log(y_{i-1} + c) - \log[\exp(X_{i-1, z_{i-1}} \beta_{z_{i-1}}) + c]$, where $X_{i-1, z_{i-1}}$ are the $(i-1)$ -th observations of the covariates for state z_{i-1} . However, when the state path is unknown, such a generalization is non-trivial. To the best of our knowledge, an AR-HMM that allows the autoregressive term to be dependent on state path and state-specific covariates has not been introduced in the literature. We develop such a model in this paper.

3. METHODS

3.1 Penalized MLE for HMMs with covariates

In a Hidden Markov Model with covariates, the observations $Y_1, \dots, Y_n$ have a natural order (e.g., observations along time points) and the transitions between latent states along the ordered observations are explicitly modeled. We again denote the random variable for state path by $Z = (Z_1, \dots, Z_n)$ and $z = (z_1, \dots, z_n)$ denotes an observed state path. Let K be the number of states, and let S be the set of K^n possible state paths of length n . We assume a stationary Markov chain with state-to-state transition probabilities $\gamma = (\gamma_{11}, \dots, \gamma_{KK})^T$, where

$\gamma_{jk} = p(Z_i = k | Z_{i-1} = j)$ for $i = 2, \dots, n$, $\sum_{k=1}^K \gamma_{jk} = 1$, $\gamma_{jk} > 0$ for all $j, k = 1, \dots, K$, and $Y_{i-1} \perp Y_i | (Z_{i-1}, Z_i)$. Then the likelihood of the observed data is

$$L_n(\Psi_H | \mathbf{X}, y) = \sum_{z \in S} \left\{ \prod_{k=1}^K [\pi_k f_k(y_1 | X_{1k}, \beta_k, \varphi_k)]^{I[z_1=k]} \prod_{i=2}^n \prod_{k=1}^K \left[f_k(y_i | X_{ik}, \beta_k, \varphi_k) \right]^{I[z_i=k]} \prod_{j=1}^K \gamma_{jk}^{I[z_{i-1}=j, z_i=k]} \right\},$$

where $\Psi_H = (\beta^T, \varphi^T, \gamma^T, \pi^T)^T$, $\beta = (\beta_1^T, \dots, \beta_K^T)^T$, and $\mathbf{X}_{n \times p}$ are the set of p covariates that may be related with the mean value of each state distribution, while the relevant covariates for each state may be a subset of the p covariates. In contrast to the notation used for the FMR, $\pi = (\pi_1, \dots, \pi_K)^T$ is now known as the set of prior probabilities of state membership for the first observation. Conditional density $f_k(y_i | X_{ik}, \beta_k, \varphi_k)$, which belongs to exponential family, is now defined as the state-specific emission density. The remaining variables are defined similarly as those for the FMR, which have been introduced in Section 2.1.

Let $l_n(\Psi_H | \mathbf{X}, y) = \log L_n(\Psi_H | \mathbf{X}, y)$ be the log likelihood. To achieve our goal in variable selection, which is to select relevant covariates pertaining to each state, we maximize the following penalized log likelihood

$$pl_n(\Psi | \mathbf{X}, y) = l_n(\Psi | \mathbf{X}, y) - \mathcal{P}(\Psi), \quad (3)$$

where Ψ is defined as $\Psi = (\beta^T, \varphi^T, \gamma^T, \pi^T, \eta^T)^T = (\Psi_H^T, \eta^T)^T$, $\eta = (\eta_1, \dots, \eta_K)^T$, and η_k is the proportion of the observations belonging to state k . $\mathcal{P}(\Psi) = \sum_{k=1}^K \eta_k \sum_{l=1}^p \rho_{\omega_k}(\beta_{lk})$ is the

total penalty to the likelihood, and $\rho_{\omega_k}(\beta_{lk})$ denotes a penalty function with tuning parameter(s) ω_k , which could be a function of the sample size n . Give the stationarity assumptions, the parameter η can be obtained from transition probability γ , however we keep η for notational simplicity.

Maximization of the penalized likelihood in (3) with respect to β balances the overall model fit, $l_n(\Psi|\mathbf{X}, y)$, and the cost of model complexity, controlled by $\mathcal{P}(\Psi)$. In this paper, we employed three penalties that represent a broad class of the available penalties.

- LASSO: $\rho_{\lambda_k}(\beta_{lk}) = \lambda_k |\beta_{lk}|$, for $\lambda_k > 0$,
- SCAD: $\rho'_{\lambda_k}(\beta_{lk}) = \lambda_k \left\{ I(|\beta_{lk}| \leq \lambda_k) + \frac{I(|\beta_{lk}| > \lambda_k)(a\lambda_k - |\beta_{lk}|)_+}{\lambda_k(a-1)} \right\}$, for $a > 2$ and $\lambda_k > 0$, where $x_+ = x$ if $x \geq 0$ and $x_+ = 0$ otherwise.
- Log Penalty: $\rho_{\lambda_k \tau_k}(\beta_{lk}) = \lambda_k \log(|\beta_{lk}| + \tau_k)$, for $\lambda_k > 0$ and $\tau_k > 0$.

The LASSO (i.e., L_1 penalty) is a convex penalty, while both SCAD and Log penalty belongs to a class of folded concave penalties [Fan and Lv, 2010]. The Log penalty can be interpreted as an Iterative Adaptive LASSO (IAL) penalty [Sun et al., 2010], which represents a class of penalties that bridge the L_0 penalty ($\rho_{\lambda_k}(\beta_{lk}) = \lambda_k I(|\beta_{lk}| > 0)$) and the L_1 penalty. The LASSO penalty has only one tuning parameter λ_k . SCAD [Fan and Li, 2001] has two regularization parameters λ_k and a . Following Fan and Li [2001], we set $a = 3.7$ for all states $k = 1, \dots, K$, and only treat λ_k as a tuning parameter. The Log penalty has two tuning parameters λ_k and τ_k .

Khalili and Chen [2007] have studied the theoretical properties of the PMLE in the content of the FMR. Specifically, they establish the conditions on penalty $p_{w_k}(\cdot)$ such that the Oracle Property can be achieved for the PMLE, which is estimated by penalized weighted least squares in the M-step of their algorithm. We extend the results of Khalili and Chen [2007] to the HMM with covariates, which requires some additional regularity conditions from Bickel et al. [1998]. Partition $\beta_k^T = (\beta_{k1}^T, \beta_{k2}^T)$ such that β_{k2} pertains to the zero effects. In addition, we partition $\Psi^T = (\Psi_1^T, \Psi_2^T)$ such that Ψ_2 contains zero parameters in the model, namely β_{k2} , $k = 1, \dots, K$. Let Ψ_0 be the true values of Ψ and $\beta_{lk,0}$ be the true regression coefficients corresponding to the l^{th} covariate in the k^{th} state. Define

$a_n = \max_{l,k} \{\rho_{\omega_k}(\beta_{lk,0}) / \sqrt{n} : \beta_{lk,0} \neq 0\}$, $b_n = \max_{l,k} \{|\rho'_{\omega_k}(\beta_{lk,0})| / \sqrt{n} : \beta_{lk,0} \neq 0\}$, and $c_n = \max_{l,k} \{|\rho''_{\omega_k}(\beta_{lk,0})| / n : \beta_{lk,0} \neq 0\}$, where $\rho'_{\omega_k}(\beta_{lk,0})$ and $\rho''_{\omega_k}(\beta_{lk,0})$ represent the first and second derivatives of $\rho_{\omega_k}(\beta_{lk})$ with respect to β_{lk} , respectively. We place the following conditions on the penalty $\rho_{\omega_k}(\beta_{lk})$:

P0: The penalty $\rho_{\omega_k}(\beta_{lk})$ is symmetric around 0, nondecreasing for β_{lk} in $(0, \infty)$ and is twice differentiable for all β_{lk} in $(0, \infty)$. $\rho_{\omega_k}(\beta_{lk})$ attains its minimum at $\beta_{lk} = 0$.

P1: As $n \rightarrow \infty$, $a_n = o_p(1 + b_n)$ and $c_n = o_p(1)$.

P2: For $N_n = \{\beta_{lk} : 0 < \beta_{lk} \leq n^{-\frac{1}{2}} \log(n)\}$, $\lim_{n \rightarrow \infty} \inf_{\beta_{lk} \in N_n} \rho'_{\omega_k}(\beta_{lk}) / \sqrt{n} = \infty$.

Corollary 1—Assume the regularity conditions apply (see Appendix Section A.1). We assume that $(Y_1, Z_1), \dots, (Y_n, Z_n)$ is a discrete-time stochastic process corresponding to the HMM with covariates such that $(Y_1, Z_1)|\mathbf{X}_1, \dots, (Y_n, Z_n)|\mathbf{X}_n$ is stationary conditional on $\mathbf{X}_i$. Then, given conditions P0–P2 and assuming the number of states K is known, we have the following conclusions:

1. Consistency: There exists a local maximizer $\hat{\Psi}$ of $pl_n(\Psi|\mathbf{X}, y)$ such that

$$\|\hat{\Psi} - \Psi\| = O_p(n^{-\frac{1}{2}}(1 + b_n)), \text{ where } \|\cdot\| \text{ represents the euclidean norm.}$$

2. Sparsity: $p(\Psi_2 = 0) \rightarrow 1$ as $n \rightarrow \infty$

3. Asymptotic Normality:

$\sqrt{n} \left\{ [I(\Psi_{01}) - \mathcal{P}''(\Psi_{01})/n] (\hat{\Psi}_1 - \Psi_{01}) + \mathcal{P}'(\Psi_{01})/n \right\} \rightarrow N(0, I(\Psi_{01}))$, where $I(\Psi_{01})$ is the subset of Fisher Information matrix for the non-zero effects, and $\mathcal{P}'(\Psi_{01})$ and $\mathcal{P}''(\Psi_{01})$ are the first and second derivatives of penalty function $\mathcal{P}(\Psi_{01})$ with respect to Ψ_{01} .

Therefore under the regularity conditions and given conditions P0–P2, the PMLE corresponding to penalty $p_{wk}(\beta_{lk})$ can achieve the Oracle Property. The proof is similar to the proofs of Theorems 2 and 3 in Khalili and Chen [2007], and we briefly described the proof in Appendix Section A.1. We note that the above theoretical properties are for the M-step estimates of the EM algorithm instead of the final estimates from the EM algorithm.

3.2 An EM + coordinate descent algorithm

In this section, we provide the details of our EM algorithm that maximizes the Penalized likelihood of a HMM. Recall that the random variable $Z = (Z_1, \dots, Z_n)$ denotes the state path. In the s -th step of the EM algorithm, the Q-function of penalized likelihood (3) is

$$\begin{aligned} Q(\Psi|\Psi^{(s)}) &= E_Z \left[p l_n(\Psi|y, \mathbf{X}, Z) | y, \mathbf{X}, \Psi^{(s)} \right] \\ &= \sum_{k=1}^K p(Z_1=k|y, \mathbf{X}, \Psi^{(s)}) \log(\pi_k) + \sum_{i=2}^n \sum_{k=1}^K \sum_{j=1}^K p(Z_{i-1}=j, Z_i=k|y, \mathbf{X}, \Psi^{(s)}) \log(\gamma_{jk}) \\ &\quad + \sum_{i=1}^n \sum_{k=1}^K p(Z_i=k|y, \mathbf{X}, \Psi^{(s)}) \log[f_k(y_i|X_{ik}, \beta_k, \varphi_k)] - \mathcal{P}(\Psi) \\ &= Q(\pi|\Psi^{(s)}) + Q(\gamma|\Psi^{(s)}) + [Q(\beta, \varphi|\Psi^{(s)}) - \mathcal{P}(\Psi)]. \end{aligned} \quad (4)$$

In the E-step, $p(Z_i=k|y, \mathbf{X}, \Psi^{(s)})$ and $p(Z_{i-1}=j, Z_i=k|y, \mathbf{X}, \Psi^{(s)})$ can be computed by the standard forward-backward algorithm, detailed in Appendix A.2. Similar to the FMR, the posterior probability $p(Z_i=k|y, \mathbf{X}, \Psi^{(s)})$ is utilized in the classification of observations. In the M-step, the Q function is separable for π , γ , and (β, φ) , and only β is penalized. Therefore, π and γ can be estimated from the unpenalized likelihood such that

$\gamma_{jk}^{(s+1)} = \left[\sum_{i=2}^n p(Z_{i-1}=j, Z_i=k|y, \mathbf{X}, \Psi^{(s)}) \right] / \left[\sum_{i=2}^n p(Z_{i-1}=j|y, \mathbf{X}, \Psi^{(s)}) \right]$, and $\pi_k^{(s+1)} = p(Z_1=k|y, \mathbf{X}, \Psi^{(s)})$. Under the assumptions of stationarity we can derive $\eta_k^{(s+1)}$ as the solution to $\eta_k^{(s+1)} \Pi(\gamma^{(s+1)}) = \eta_k^{(s+1)}$, where $\Pi(\gamma^{(s+1)})$ is the $K \times K$ transition probability matrix based on $\gamma^{(s+1)}$. For simplicity we estimate η_k such that $\eta_k^{(s+1)} = \sum_{i=1}^n p(Z_i=k|y, \mathbf{X}, \Psi^{(s)})/n$. This estimate works well in our simulations.

$Q(\beta, \varphi|\Psi^{(s)})$ can be decomposed into K components, one for each state. Therefore we can maximize the last term of (4) with respect to β_k and φ_k separately for each state k . One approach is to alternately estimate β_k and φ_k until convergence [Hilbe, 2011]. However, this approach is computationally intensive and we adopt a one-step update in our algorithm.

Specifically, we perform a conditional maximization to obtain $\beta_k^{(s+1)}$ given $\varphi_k^{(s)}$ using penalized Iteratively Reweighted Least Squares (IRLS) followed by an conditional

maximization to obtain $\varphi_k^{(s+1)}$ given $\beta_k^{(s+1)}$. Our algorithm can be considered as an ECM algorithm [Meng and Rubin, 1993] where we perform conditional maximization of β_k and φ_k . In contrast to alternately estimating β_k and φ_k until convergence for each M-step, our one-step update of β_k and φ_k leads to more iterations in the ECM algorithm, but overall less

computational time. The details of this ECM algorithm are presented in the Supplementary Materials Section 1.1.

Here we briefly describe a key part of this algorithm, the penalized IRLS to estimate $\beta_k^{(s+1)}$. Employing a canonical link function, we can derive the following objective function of the penalized IRLS:

$$\mathcal{Q}_k(\beta_k | \Psi^{(s)}) = \frac{1}{2} \sum_{i=1}^n \zeta_{ik}^{(s)} \left[v_{ik}^{(s)} (q_{ik} - X_{ik} \beta_k)^2 \right] + \eta_k^{(s)} \sum_{l=1}^{p_k} \rho_{\omega_k}(\beta_{lk}), \quad (5)$$

where

$\zeta_{ik}^{(s)} = p(Z_i = k | y, \mathbf{X}, \Psi^{(s)})$, $\mu_{ik}^{(s)} = E(y_i | Z_i = k, X_{ik}, \Psi^{(s)})$, $v_{ik}^{(s)} = \text{Var}(y_i | Z_i = k, X_{ik}, \Psi^{(s)})$, and $q_{ik}^{(s)} = X_{ik} \beta_k^{(s)} + (y_i - \mu_{ik}^{(s)}) / v_{ik}^{(s)}$. We minimize the above objective function by a coordinate descent algorithm. Prior to minimization, we standardize the columns of $\mathbf{X}$ to be mean 0 and variance 1, and we transform the final estimates of $\hat{\beta}_k$ back to their unstandardized values following convergence of the coordinate descent algorithm.

To select tuning parameters, we follow the procedure similar to Khalili and Chen [2007] where we first obtain the MLE under the full model Ψ_{full} on the data. We then select the optimal set of tuning parameters for each state individually, while fixing the parameters of all other states at their full model MLEs. This procedure significantly reduces the computational cost in tuning parameter selection when a large number of states exist. For each state, we select tuning parameters by minimizing BIC.

3.3 Autoregressive Hidden Markov Model with Covariates (AR-HMM)

We extend the HMM with covariates described in the Section 3.1 to allow dependence between the observations conditional on the hidden states. Given underlying states, we assume that there is AR(1) dependence between Y_i and Y_{i-1} conditional on (Z_{i-1}, Z_i) such that $f(y_i | y_{i-1}, \dots, y_1, Z_{i-1} = j, Z_i = k, \mathbf{X}, \beta, \phi) = f_{jk}(y_i | X_{ik}, \beta_k, \phi_k, \nu_k, r_{i-1,j})$, where the subscript jk in f_{jk} indicates $Z_{i-1} = j, Z_i = k$ and

$$r_{i-1,j} = \begin{cases} 0 & \text{if } i=1 \\ \log(y_{i-1} + 1) - \log[\exp(X_{i-1,j} \beta_j) + 1] & \text{if } i > 1. \end{cases} \quad (6)$$

If the underlying state is k for the i -th observation, then $E(y_i | Z_i = k, Z_{i-1} = j, X_{ik}, \beta_k, \phi_k, \nu_k, r_{i-1,j}) = \mu_{ikj}$ with link function $h(\mu_{ikj}) = X_{ik} \beta_k + \nu_k r_{i-1,j}$ where ν_k is the set of AR coefficient for state k . Then the AR-HMM complete data likelihood given some state path z is

$$L(\Psi_A | \mathbf{X}, y, z) = \prod_{k=1}^K [\pi_k f_k(y_1 | X_{1k}, \beta_k, \phi_k)]^{I[z_1=k]} \times \prod_{i=2}^n \prod_{k=1}^K \prod_{j=1}^K [\gamma_{jk} f_{jk}(y_i | X_{ik}, \beta_k, \phi_k, \nu_k, r_{i-1,j})]^{I[z_{i-1}=j, z_i=k]},$$

where $\Psi_A = (\beta^T, \phi^T, \gamma^T, \pi^T, \nu^T)^T$ and $\beta = (\beta_1^T, \dots, \beta_K^T)^T$.

We develop an EM algorithm inspired by Ibrahim [1990] to obtain the MLE of the AR-HMM. We can show that the Q-function is

$$\begin{aligned}
Q(\Psi_A | \Psi_A^{(s)}) &= E_Z \left\{ \log [L(\Psi_A | y, \mathbf{X}, z)] | \Psi_A^{(s)}, y, \mathbf{X} \right\} \\
&= \sum_{k=1}^K p(Z_1=k | y, \mathbf{X}, \Psi_A^{(s)}) \log(\pi_k) + \sum_{i=2}^n \sum_{k=1}^K \sum_{j=1}^K p(Z_{i-1}=j, Z_i=k | y, \mathbf{X}, \Psi_A^{(s)}) \log(\gamma_{jk}) \\
&\quad + \sum_{i=1}^n \sum_{k=1}^K \sum_{j=1}^K p(Z_{i-1}=j, Z_i=k | y, \mathbf{X}, \Psi_A^{(s)}) \log[f_{jk}(y_i | X_{ik}, \beta_k, \phi_k, \nu_k, r_{i-1,j})].
\end{aligned}$$

Direct maximization of the above Q function is computationally difficult because β is also present in $r_{i-1,j}$. We adopt an approximation to fix $r_{i-1,j}$ at $r_{i-1,j}^{(s)}$, which is the value of $r_{i-1,j}$ at step s given $\beta^{(s)}$. This approximation significantly improves the computational efficiency of our algorithm, which is very important for the analysis of DAE-seq data with tens of thousands of observations. Later simulation results show that this approximation does not lead to any bias of MLE. At time point $i = 1$, the likelihood that the current state is k is weighted by $p(Z_1=k | y, \mathbf{X}, \Psi_A^{(s)})$, and for $i > 1$ the likelihood that the current state is k and previous state is j is weighted by $p(Z_{i-1}=j, Z_i=k | y, \mathbf{X}, \Psi_A^{(s)})$. We can derive these quantities from the forward and backward probabilities, see the Appendix Section A.2 for details.

Given the weights from the E-step, we obtain the MLE of Ψ_A in the M-step. Since the Q-function can be separated into three sets of parameters π , γ , and $(\beta^T, \phi^T, \nu^T)^T$, we can estimate each set of parameters separately. First,

$$\pi_k^{(s+1)} = p(Z_1=k | y, \mathbf{X}, \Psi_A^{(s)}), \gamma_{jk}^{(s+1)} = \sum_{i=2}^n \tau_{ijk} / \left[\sum_{i=2}^n \sum_{k=1}^K \tau_{ijk} \right] \text{ where}$$

$\tau_{ijk} = p(Z_{i-1}=j, Z_i=k | y, \mathbf{X}, \Psi_A^{(s)})$. We estimate β_k , ϕ_k , and ν_k for each state k using the following augmented regression to account for missing data due to the AR component in the model. Following Ibrahim [1990], let $\tilde{y}$ be the augmented version of y by repeating each y_i

K times. In other words, $\tilde{y} = (\underbrace{y_1, \dots, y_1}_K, \underbrace{y_2, \dots, y_2}_{K-}, \dots, \underbrace{y_n, \dots, y_n}_K)^T$. Let $\tilde{\mathbf{X}}$ be the augmented version of $\mathbf{X}$. The dimension of $\tilde{\mathbf{X}}$ is $nK \times (p+1)$. The first p columns of $\tilde{\mathbf{X}}$ is constructed by repeating each row of $\mathbf{X}$ K times. Let $\mathbf{r}_i = (r_{i-1,1}^{(s)}, \dots, r_{i-1,K}^{(s)})^T \in \mathbb{R}_{K \times 1}$, where $r_{i-1,j}^{(s)}$ is defined in (6) given $\beta^{(s)}$, and we set the $(p+1)$ th column of $\tilde{\mathbf{X}}$ as

$\mathbf{r} = (\mathbf{r}_1^T, \dots, \mathbf{r}_n^T)^T \in \mathbb{R}_{nK \times 1}$. We construct $\tilde{\mathbf{X}}_k \in \mathbb{R}_{nK \times (p_k+1)}$ by extracting the p_k columns of $\tilde{\mathbf{X}}$ corresponding to the p_k covariates for state k and the $(p+1)$ th column of $\tilde{\mathbf{X}}$. Then the parameters β_k , ϕ_k and ν_k can be estimated by a weighted generalized linear regression of $\tilde{y}$ on $\tilde{\mathbf{X}}_k$ with the weights $\mathbf{w}_k = (\mathbf{w}_{1k}^T, \dots, \mathbf{w}_{nk}^T)^T$, $\mathbf{w}_{ik} = (w_{i1k}, \dots, w_{iKk})^T$ for $i = 1, \dots, n$, and $w_{ijk} = p(Z_{i-1}=j, Z_i=k | y, \mathbf{X}, \Psi_A^{(s)})$, for $j = 1, \dots, K$. This is equivalent to complete data maximum likelihood estimation where the missing data is “filled in” with a set of weighted values spanning the range of the discrete missing covariate, in this case $r_{i-1,j}$. The approach in Lystig and Hughes [2002] allows for computationally efficient and exact computation of the observed information matrix using a modified forward-backward algorithm.

The penalization procedure described in Section 3.1 extends to the AR-HMM by simply replacing $\mathbf{X}$ with $\tilde{\mathbf{X}}$, y with $\tilde{y}$, and utilizing weights w_{ijk} from the AR-HMM E-step. This procedure is similar to penalized estimation with missing data [Garcia et al., 2010], but the AR part complicates the asymptotic theory. We expect the oracle properties similar to Corollary 1 to hold, but a careful theoretical study is beyond the scope of this paper.

4. SIMULATION STUDIES

In this section, we evaluate the performance of the proposed variable selection procedure in simulated data. We also compare the classification performance and accuracy of the AR-HMM relative to other methods in simulated DAE-seq datasets.

4.1 Simulation Setup

Simulated datasets of a two-state Poisson AR-HMM with first order dependence in states and AR(1) dependence in observations were generated in the following manner. We simulated data to mimic window read counts from a CTCF (CCCTC-binding factor) ChIP-seq dataset and a H3K36me3 (Trimethylation of Lys36 in histone H3) histone modification ChIP-seq data set. In the CTCF dataset, enrichment regions are short, rare, and contain much higher signals relative to background. This enrichment pattern is typical for most transcription factor binding sites. In contrast, enrichment regions from H3K36me3 dataset are typically broader, more abundant, and contain weaker signals. We simulated window read counts corresponding to these two DAE-seq data types to represent a wide range of enrichment patterns found in real data analysis.

For both simulated data types, we set $\gamma_{11} = 0.9$ and $\gamma_{12} = 0.1$, which correspond to the background-to-background and background-to-enrichment state transition probability, respectively. For simulated CTCF ChIP-seq data, we set $\gamma_{21} = 0.9$ and $\gamma_{22} = 0.1$, corresponding to the enrichment-to-background and enrichment-to-enrichment transition probability, respectively. For simulated histone modification data we set $\gamma_{21} = 0.1$ and $\gamma_{22} = 0.9$. This simulation setup results in rare transitions from background to enriched regions in both data types, shorter regions of enrichment in the CTCF data, and broader regions of enrichment in the histone modification data.

The underlying state path z was simulated using the transition probabilities corresponding to each simulated data type. For each window i ($i = 1, \dots, n$), we simulated a set of p covariates $X_i = (x_{i1}, \dots, x_{ip})$ as uniform (0,1) random variables to generate covariate matrix $\mathbf{X}_{n \times p}$. We utilized the same $\mathbf{X}$ to simulate window read counts corresponding to either of the $K = 2$ states. In each data type, the relative strength of the signal in each state can be tuned by modifying β_1 and β_2 appropriately. Then, given X , z and the selected model parameters, we recursively simulated window read counts $y_1, \dots, y_n$ for each simulation case in the AR-HMM using the following procedure:

1. for $i = 1$, $y_1 \sim \text{Pois}(\exp\{X_1\beta_k\})$ if $z_1 = k$,
2. for $i > 1$, $y_i \sim \text{Pois}(\exp\{X_i\beta_k + \nu_k [\log(y_{i-1} + 1) - \log(\exp(X_{i-1}\beta_j) + 1)]\})$ if $z_i = k$ and $z_{i-1} = j$.

To simulate window read counts from HMM with covariates, we simply followed the above procedure except we set $\nu_1 = \nu_2 = 0$.

4.2 Variable Selection in Hidden Markov Models with Covariates

In the following simulation studies, we evaluate our variable selection method in the context of the HMM with covariates using small ($n = 200$) or large ($n = 10000$) sample size and low ($p = 5$) or high ($p = 100$) dimension. We employ three penalties: the LASSO, SCAD, and Log penalties. The true parameter values corresponding to the first 4 covariates per component are listed in Table 1, and all the other coefficients are set to be 0. We utilize the same set of regression parameters to simulate both the CTCF and histone modification-style datasets so that we can directly compare the effect of relative state frequencies on variable selection. Variable selection performance is measured by the number of true discoveries (TDs) and the number of false discoveries (FDs). Specifically, among all the covariates

selected by a variable selection method, a TD is a covariate that has (true) non-zero coefficient and a FD is a covariate that has (true) zero coefficient. These numbers of TDs/FDs are averaged across 100 simulations for each simulation situation.

Table 2 lists the simulation results for each simulation condition. Overall, the number of TDs increases and the number of FDs decreases as the sample size n increases. We observe that of all the penalties, the LASSO has the worst variable selection performance and greatest bias in the estimated values for the true non-zero parameters. This is in line with the results from [Fan and Li, 2001, Khalili and Chen, 2007], since the LASSO cannot satisfy all of the penalty conditions P0–P2 for the Oracle Property. As a result, it cannot simultaneously achieve sparsity and unbiased estimation of the true non-zero coefficients as $n \rightarrow \infty$. The Log and SCAD penalties, however, satisfy these conditions, and have substantially better performance than the LASSO. These results provide empirical support for the Oracle Property of Corollary 1.

In this simulation, approximately 10% of simulated windows in the simulated CTCF-style data are from the enrichment state, in contrast, ~50% of simulated windows in the histone modification-style data are from enrichment state. Comparing the performance across the two simulated data types, the variable selection performance and parameter estimation accuracies decrease when the relative state frequencies are unbalanced, such as in the simulated CTCF-style data. However given the large sample sizes that are typical in DAE-seq datasets, the effect of this imbalance will be limited.

4.3 AR-HMM

For each simulated data type (e.g., CTCF or histone modification), we simulated 1000 datasets of $n = 10,000$ observations each from a two-state Poisson AR-HMM with first order dependence in states and AR(1) dependence in observations. The mean value of each state-specific emission distribution is a function of two covariates plus an intercept. To simulate CTCF-style data with higher levels of signal in the enrichment state relative to background, we set $\beta_1 = (\beta_{01}, \beta_{11}, \beta_{21}) = (0, 1, 1)$ and $\beta_2 = (\beta_{02}, \beta_{12}, \beta_{22}) = (1.5, 2, 2)$. In the histone modification-style data, we set $\beta_1 = (\beta_{01}, \beta_{11}, \beta_{21}) = (0, 1, 1)$ and $\beta_2 = (\beta_{02}, \beta_{12}, \beta_{22}) = (0.5, 2, 2)$ to simulate weaker signals in the enrichment state. Within each simulated data type, we allowed ν_2 to be either 0.2 or 0.8 (weak or strong auto-correlation) and we fixed ν_1 to be 0.2 to mimic the observed low dependence between windows in background (Figure 1). For each simulation case, we compared the parameter estimates and classification performance of the AR-HMM with those from the FMR and the HMM with covariates. The AR-HMM estimates are accurate regardless of the values for ν_2 or simulated data type (Table 3), suggesting that the AR-HMM estimation procedure is robust over a range of conditions. In contrast, the estimates from the HMM and FMR tend to be biased in each simulation setting. The magnitude of the bias increases as the value of ν_2 increases. This bias however is larger in the simulated histone modification-style data. In the simulated CTCF-style data, the parameter estimates for the FMR and HMM are very similar (differences are on the order of 10^{-5}). This is due to the fact that the majority of transitions in the CTCF-style data are background-to-background, and that the enrichment regions are relatively easy to discern by each method due to their strong signals. Therefore, accounting for dependence in states alone in the HMM does not yield better accuracy in parameter estimates relative to the FMR.

Our main interest however is the performance of each method to distinguish enriched and background regions. We evaluated such classification performance by ROC curves (Figure 2). In the simulated CTCF ChIP-seq data with $\nu_2 = 0.2$ (Figure 2A), all methods perform similarly. This is expected, as in CTCF ChIP-seq the strong and sharp signals in enrichment regions allow for adequate detection of enrichment even in the absence of any covariate

information [Rashid et al., 2011]. When the dependence between observations in the enrichment state increases from $\nu_2 = 0.2$ to $\nu_2 = 0.8$, the AR-HMM performs slightly better than other methods (Figure 2B).

However, in the simulated histone modification-style data, the AR-HMM performs much better relative to other methods. When the dependence between observations from the enrichment state is low ($\nu_2 = 0.2$), both the HMM and AR-HMM perform much better than the FMR (Figure 2C). This is because the FMR cannot account for the more prevalent enrichment-enrichment transitions between windows, which can aid the detection of regions containing weaker enrichment signals. When the dependence between observations from the enrichment state is high ($\nu_2 = 0.8$), the AR-HMM performs much better than both the HMM and FMR (Figure 2D).

We also observe that under model misspecification, where there is no correlation in the underlying states and observations given the states, the AR-HMM performs similarly to the correct model: the FMR. For example, using CTCF style data simulated under FMR assumptions, we find that the parameter estimates for the AR-HMM, HMM, and FMR are almost the same, and the estimates for autoregressive parameters ν_1 and ν_2 are close to 0 (Supplementary Table 1). Therefore these methods have the same performance to identify enriched regions.

4.4 Variable Selection in the AR-HMM

Next we demonstrate the performance of variable selection method in AR-HMM. We generated simulated data sets similar to Section 4.2, except that we allow for dependence between simulated observations given the states by setting $\nu_1 = \nu_2 = 0.4$. Similar to the results from Section 4.2, variable selection performance improves and estimation bias drops as the sample size increases (Table 4). The estimation accuracy of ν_k , which we do not penalize, also increases with sample size. For CTCF style data, the variable selection performance is worse in the high-dimensional low sample size case ($p=100$ and $n=200$), owing to the small number of samples in the enrichment state in the simulation (approximately 20). In real data analyses we typically observe sample sizes much larger than $n = 200$ so we do not expect this to be an issue. Other conclusions with respect to data type and penalties are similar to what are observed in the case of HMM with covariates. Empirically, these results demonstrate that variable selection performance is adequate in the AR-HMM and the PMLEs in this context share similar properties to those in the HMM with covariates.

5. APPLICATION TO HUMAN GM12878 CTCF AND H3K36ME3 CHIP-SEQ DATASETS

5.1 Data preparation and model selection

We benchmarked the performance of the FMR, HMM, and the AR-HMM in two ChIP-seq datasets in terms of their ability to identify biologically relevant signals. These datasets were obtained from the ENCODE project [Bernstein et al., 2012] and included a human GM12878 CTCF ChIP-seq dataset (UT-Austin, Replicate 3) and a human GM12878 H3K36me3 ChIP-seq dataset (Broad Histone, Replicates 1 and 2). In the CTCF ChIP-seq data, we checked whether the significant regions called by each method overlapped with CTCF binding motifs, which are conserved DNA sequences that the CTCF transcription factor preferentially binds to [Kim et al., 2007]. H3K36me3 histone modifications are deposited broadly across gene bodies during transcription [Barski et al., 2007], and thus we benchmarked the enriched regions of H3K36me3 histone modifications by their overlap with gene bodies.

In each dataset, non-overlapping 250bp windows from Human Chromosome 22 were utilized to tabulate window read counts. Covariate information and read counts for each window were tabulated in the manner detailed in Rashid et al. [2011]. In this analysis we considered the covariates GC-content, mappability, and window read counts from a matching input control. We then applied the two-state Negative Binomial AR-HMM, two-state Negative Binomial HMM, and two-component Negative Binomial FMR model to each dataset. For each method, the mean value of each state distribution was modeled with some covariates using a log link function.

Each of these methods can calculate the posterior probability for each window belonging to background. Denote the posterior probability that the i -th window belongs to background by κ_i , ($i = 1, \dots, n$). Such κ_i 's are also referred to as local FDRs [Efron et al., 2001] for detecting enriched regions. For a cutoff of posterior probability α , the total FDR is $\sum_{i=1}^n [\kappa_i I(\kappa_i \leq \alpha)] / \sum_{i=1}^n [I(\kappa_i \leq \alpha)]$. We chose a posterior probability cutoff by controlling FDR. Adjacent windows meeting a given FDR threshold were merged together into a single region, and multiple performance metrics were calculated for the set of enriched regions identified by each method.

It is not known *a priori* which set of covariates should be used to model the mean of each state-specific emission distribution. Therefore, we employed the proposed variable selection procedure to determine the best model for HMM and AR-HMM in each dataset. The full model includes an intercept (fixed), the main effects of mappability, GC content, and input control, as well as their two-way and three-way interactions. In the AR-HMM model, we included the autoregressive covariate from (2) but did not subject it to penalization. Given the simulation results from Tables 2 and 4, we used the SCAD penalty in our real data application. Including the main effects and interactions, there are 7 covariates for the mean model of each state, hence 128 possible models per state and 16384 models for two states. Therefore all-subset selection is infeasible even in this relatively simple situation. In regression studies involving interactions, a reasonable constraint is that higher order interactions are included in the model if and only if all the corresponding main effects and lower order interactions are also included in the model. We did not implement this constraint because of computational challenge and because these covariates and their interactions were not of biological interest. The benefit of variable selection of these covariates was to provide an automatic procedure for model fitting. An example of selecting biological meaningful factors is presented in the next section.

We find that in each dataset, the model selected for the AR-HMM has much better fit than the model selected for the HMM in terms of BIC (Table 5). In addition, the AR-HMM estimates for ν_2 in both datasets are large, suggesting that strong dependence exists between window read counts in enrichment regions. In background regions, this dependence is much weaker, where the estimate for ν_1 are 0.283 and 0.163 for the CTCF and H3K36me3 ChIP-seq datasets, respectively. We also observe that each selected model of background state includes the three-way interaction of GC content, mappability, and input control ($\beta_{123,1}$), suggesting a strong synergistic relationship in their effects on background signals.

5.2 Performance comparison for CTCF ChIP-seq

Given these selected models, we first examined the classification performance of the FMR, HMM, and the AR-HMM in the CTCF ChIP-seq data across FDR cutoffs. For the FMR, we utilized the model selected for the HMM for all of the subsequent analyses. In the CTCF ChIP-seq data, both the AR-HMM and HMM call less enriched regions than the FMR (Figure 3A). A slightly higher proportion of the enriched regions called by the AR-HMM or

the HMM overlap with CTCF binding sites (Figure 3C), which is partly due to the enriched regions called by both methods tending to be longer (Figure 3B, F).

Next we compared the performances of different methods using ROC curves while defining a true discovery as the window/region that overlaps a CTCF motif. The three methods perform similarly (Supplementary Figure 1), and the FMR performs slightly better in terms of number of windows overlapping CTCF sites. This is because AR-HMM and HMM tends to call longer regions that cover more windows than FMR (Figure 3F). The majority of significant regions called by each method overlap those called by other methods (Table 6), and the maximum signals of the significant regions that are called uniquely by each method are much greater than the background (Figure 3D), which suggest that none of methods call many false positives. Therefore we conclude that the three methods perform similarly in the CTCF data. This is expected given the simulation results and the fact that CTCF data have strong and easily discernible enriched regions. We notice, however, that the enrichment-enrichment transition probability (γ_{22}) of the HMM and AR-HMM (Table 7) is relatively large, suggesting that some regions of CTCF ChIP-seq enrichment may span more than a single window in real data. Examples are shown in Figure 3E,F. Finally, we also examined the performance of two popular existing methods F-seq [Boyle et al., 2008] and MACS [Zhang et al., 2008a]. Similar to results from Rashid et al. [2011], we found all methods perform similarly for CTCF ChIP-seq data (Supplementary Figure 2A).

5.3 Performance comparison for the H3K36me3 ChIP-seq data

In contrast to the CTCF ChIP-seq data, the enrichment regions in H3K36me3 ChIP-seq are much broader and have relatively weaker signals. We benchmark the enriched region calls by their coverage of gene bodies rather than whether an enriched region has overlap with any portion of a gene body. We observe a significant improvement in performance of the AR-HMM relative to the HMM or FMR. For example, enriched regions called by the AR-HMM (“AR-HMM calls” for short) generally span greater lengths of gene bodies (Figure 4A) and each AR-HMM call tend to be longer than HMM calls or FMR calls (Figure 4B). Although the FMR calls overlap more genes than the AR-HMM calls (Figure 4C), the AR-HMM calls cover a greater average proportion of the overlapped gene bodies (Figure 4D). ROC curves confirm that the AR-HMM and HMM have acceptable specificity (Supplementary Figure 1C). The enriched windows identified by AR-HMM and HMM cover ~20% of Chr22. About 90% of these enriched regions overlap with a gene body, which is much higher than expected by chance considering less than 40% of Chr22 are covered by gene bodies (including both intronic and exonic regions). The performance difference is most apparent in regions where the enrichment signal is relatively weak. For example, in the regions shown in Figure 4E–F, the AR-HMM and HMM tends to classify consecutive windows into enriched regions while the FMR calls are much more sporadic. As a result, the number of regions called by FMR is much greater than those from AR-HMM and HMM (Supplementary Figure 1D), but overall these regions covered lower portions of gene bodies (Supplementary Figure 1C).

In addition, we applied MACS and F-seq to detect enriched regions in the H3K36me3 ChIP-seq data. We find that our AR-HMM and HMM perform significantly better in terms of sensitivity and specificity of gene body coverage relative to these methods (Supplementary Figure 2B). Therefore, based on the above results of real data analysis and simulations, we conclude that accounting for multiple sources of dependence in the observations may significantly improve the performance of detecting enriched regions in epigenetic datasets.

All the previous results are based on non-overlapping windows. Many analyses utilize overlapping windows to account for possible window “boundary effects”, where regions of elevated signal may be split by a window’s boundary [Zhang et al., 2008b, Ji et al., 2008]. In

such situations, adjacent windows have stronger AR correlations because they are partially overlapped, and thus we would expect the AR-HMM to have a greater advantage over the HMM. To illustrate this point, we performed real data analysis in our H3K36me3 dataset using overlapping windows (250bp windows with 125bp overlap). Supplementary Figure 3A confirms that the advantage of AR-HMM is larger when using overlapping windows than non-overlapping windows. In fact, HMM calls include more false positives when studying read counts of overlapping windows, as it cannot distinguish correlation due to underlying states dependence or due to AR dependence (Supplementary Figure 3B–D).

6. THE RELATION BETWEEN HISTONE MODIFICATION H3K36ME3 AND TRANSCRIPTION FACTOR OCCUPANCY

The functional role of histone modification H3K36me3 has attracted a great amount of research interest. It has been shown that H3K36me3 is involved in the elongation phase of transcription [Li et al., 2007], leukaemogenesis [Wang et al., 2007], mRNA splicing [Kolasinska-Zwierz et al., 2009], and DNA mismatch repair [Li et al., 2013]. In earlier sections of this paper we sought to classify genomic regions as either H3K36me3-enriched or background. However, the magnitude of ChIP-seq signals within enriched or background regions itself is also biologically meaningful since it reflects the proportion of cells having a H3K36me3 mark at that location, among a large population of cells. Furthermore, the genome-wide variability of these signals in enriched/background regions may be associated with a subset of biological factors. However, no method currently exists to efficiently discover *state-specific* relationships between DAE-seq signals and a large number of biological factors.

In this section, we use our variable selection procedure and the ChIP-seq data from the ENCODE (Encyclopedia of DNA Elements) project [Thurman et al., 2012] to study the state-specific relationship between H3K36me3 signals and the DNA binding signals of 40 transcription factors (TFs) in either H3K36me3-enriched regions or background regions. We study this relationship in two ways. First, we study the relationship between H3K36me3 signal and TF binding within the same window. Then, we assess the association between TF binding in promoter regions and H3K36me3 signals in downstream genes. The former study examines state-specific relationships between H3K36me3 and local TF binding, while the later directly examines promoter-driven regulation of H3K36me3 signal across gene bodies.

The ENCODE ChIP-seq data utilized in this study were all generated from the K562 cell line, which is a myelogenous leukemia line derived from a 53 year old female CML (chronic myelogenous leukemia) patient [Lozzio and Lozzio, 1975]. We downloaded ChIP-seq data of H3K36me3 and 40 transcription factors including RNA polymerase II (Pol2) from the UCSC Genome Browser (See Supplementary Table 2 for the list of bam files). All downloaded files correspond to untreated samples with reads mapped to human genome build hg19. The H3K36me3 data have ~25 million reads. To normalize for read-depth differences, we randomly down-sampled each TF dataset to approximately 10 million reads. Then for each dataset we counted the number of reads in 250 bp non-overlapping windows in chromosome 19 similar to Rashid et al. [2011], resulting in approximately $n = 220,000$ windows per sample. For each window i ($i = 1, \dots, n$), we had window read counts y_i corresponding to the H3K36me3 data and window read counts X_{il} , $l = 1, \dots, 40$ from each of the $p = 40$ TF ChIP-seq datasets.

Utilizing our penalized AR-HMM with log penalty, we first seek to select covariates related with y_i using TF binding signals within the same window. There are 47 covariates that include variables for the 40 TFs and 7 possible confounding effects: mappability, GC content, input control, and their 2-way or 3-way interactions. In the following discussion, we

omit the variable selection results for the confounding effects since they are not of biological interest. To avoid the over-dispersed nature of count data, we test three transformations of the TF count data: $\log(X_{il} + 1)$, $I(X_{il} > q_{X_{il},90})$, and $I(X_{il} > q_{X_{il},95})$, where $I()$ is an indicator function and $q_{X_{il},\alpha}$ indicates the α percentile of X_{il} . The thresholding of the window read counts of a TF serves to be a binary approximation of a TF binding event. The variable selection and parameter estimation results for these three transformations are similar (Supplementary Figure 4, Supplementary Table 3) and thus we only summarize the results from transformation $I(X_{il} > q_{X_{il},95})$.

As shown in Figure 5, the TFs selected in the background state and the enriched state have some similarities. In both states, the TF with strongest association with H3K36me3 is RNA Polymerase II (Pol2), which is expected given the involvement of H3K36me3 in transcriptional elongation. The TF with the next strongest association with H3K36me3 in both states is ZBTB7, which has been shown to interact with histone deacetylase-1 [Liu et al., 2004], and thus our results imply the possibility of interplay between histone methylation and acetylation. ZBTB7 is also related with leukemia, where it is also known as Leukemia/lymphoma-related factor. Given H3K36me3's known association with leukemia, it would be interesting to study whether the association between H3K36me3 and ZBTB7 is specific in leukemia cell lines. In addition, BRG1 binding is positively related with H3K36me3 signals in the background but not the enrichment state, in line with its known role of the selective remodeling of chromatin structure outside of genes to aid in the recruitment of transcription factors [Roberts and Orkin, 2004]. Other factors exhibit weaker effects which may suggest less frequent interactions associated with local H3K36me3 deposition or background signal.

We would like to clarify that the above analysis is different from more commonly used analyses utilized to examine the marginal correlation between H3K36me3 and an individual TF in three aspects. First, we assess associations within H3K36me3-enriched and background regions separately, instead of performing genome-wide association. Second, these associations are conditioned on the signals of all other TFs present in the model, which may be different from marginal associations. For example, a TF may modify H3K36me3 through Pol2 regulation, and thus marginally associated with H3K36me3. However such association may be attenuated given Pol2 signals. Third, we examine the association of H3K36me3 signals and TF bindings within the same window whereas previous studies sometime examine TF bindings at gene promoters.

While it is not unreasonable to expect that certain DNA-protein binding events may directly affect local H3K36me3 deposition or background signal, another biologically interesting situation is to examine the association between TF binding at promoters and H3K36me3 at downstream genes. Since H3K36me3 typically covers gene bodies of actively transcribed genes, in this setup H3K36me3-enriched and background regions would arise from those genes with high and low transcriptional activity vs. those with no or low transcriptional activity, respectively. We have conducted such an analysis to focus on H3K36me3 signals along gene bodies, adjusting for confounding factors as in the previous study but now defining the TF covariate for an entire gene as a binary variable indicating promoter region binding of that particular TF for that particular gene (See supplementary materials section 1.3 for the details of data preparation). We applied our penalized AR-HMM with log penalty to this data. Since two adjacent genes may be far apart in the genome, we reset the autoregressive covariate to be 0 at the beginning of each new gene in the data matrix to avoid unjustified autoregressive effects. Similar to previous study, we found H3K36me3 is negatively associated with ZBTB7 binding, and positively associated with Pol2 in both H3K36me3-enriched and background regions (Supplementary Figure 5). In contrast, some TFs show different effects in these two analyses. For example, in previous analysis, cMYC

binding is not associated with H3K36me3 in enriched regions and is negatively correlated with H3K36me3 in background regions (Figure 5). However cMYC binding in promoters show strong positive effects on H3K36me3 in both H3K36me3-enriched and background regions, in line with its role as a transcriptional activator [Felsher et al., 2000].

In summary, we find the occupancy of multiple TFs are associated with H3K36me3 signatures and such associations may vary between H3K36me3-enriched regions and background regions. The functions of these TFs, together with the involvement of H3K36me3 in cancer-related processes imply interesting connections between chromatin modification and tumorigenesis, a theme that is attracting increasing interest recently [Suvà et al., 2013].

7. CONCLUSIONS AND DISCUSSIONS

We have proposed and implemented two novel strategies for DAE-seq data analysis: to account for dependency of DAE-seq data from adjacent genomic loci using HMM/AR-HMM with covariates, and to conduct variable selection in the setup of HMM or AR-HMM. Our simulation and real data analysis results suggest an existing approach of Finite Mixture Regression (FMR) model is sufficient for DAE-seq data where signal-to-noise ratio is high and the enriched regions are short. When the enriched regions are longer, HMM and AR-HMM show advantages. When there are autocorrelations between adjacent windows (which is a natural consequence of using overlapping sliding windows) given hidden state, AR-HMM performs better than the other methods. We show that even if the true model is FMR, both HMM and AR-HMM perform well. In addition, some DAE-seq data may have a mixture of two types of patterns: sharp peaks and segmental low-signal enrichments. Therefore applying AR-HMM is much more convenient for real data analysis. We applied our variable selection method in a chromosome-wide analysis and a gene-centered analysis. This type of study can be conducted in genome-wide scale or more focused regions such as the genes belonging to the same pathway. The response variable can be other quantitative features such as open chromatin regions captured by DNase-seq [Thurman et al., 2012].

We have implemented our methods in an R package that can be downloaded from <http://code.google.com/p/hmmcov/>. Our software implementation is computationally efficient. For example, in our real data analysis for CTCF or H3K36me3, to analyze ~140,000 non-overlapping windows spanning Chr22, it takes less than 120/180 seconds for HMM and AR-HMM, respectively; and to analyze ~280,000 overlapping windows spanning Chr22, it takes less than 220/540 seconds for HMM and AR-HMM, respectively. For the real data analysis of H3K36me3 signals versus 47 covariates (40 TFs + 7 confounding factors) at Chr19, the total computational time is less than 4 hours with 25 tuning parameter combinations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work is supported in part by NIH grant R01 GM070335, Dr. Wei Sun's research is also supported in part by R03 CA167684.

References

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K, et al. High-resolution profiling of histone methylations in the human genome. *Cell*. 2007; 129(4):823–837. [PubMed: 17512414]

- Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012; 489(7414):57. [PubMed: 22955616]
- Bickel PJ, Ritov Y, Ryden T. Asymptotic normality of the maximum-likelihood estimator for general hidden markov models. *The Annals of Statistics*. 1998; 26(4):1614–1635.
- Boyle, Alan P.; Guinney, Justin; Crawford, Gregory E.; Furey, Terrence S. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics*. 2008; 24(21):2537–2538. [PubMed: 18784119]
- Cox DR, Gudmundsson G, Lindgren G, Bondesson L, Harsaae E, Laake P, Juselius K, Lauritzen SL. Statistical analysis of time series: Some recent developments [with discussion and reply]. *Scandinavian Journal of Statistics*. 1981:93–115.
- Davis, Richard A.; Dunsmuir, William TM.; Streett, Sarah B. Observation-driven models for poisson counts. *Biometrika*. 2003; 90(4):777–790.
- Efron, Bradley; Tibshirani, Robert; Storey, John D.; Tusher, Virginia. Empirical bayes analysis of a microarray experiment. *Journal of the American statistical association*. 2001; 96(456):1151–1160.
- Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*. 2001; 96(456):1348–1360.
- Fan J, Lv J. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*. 2010; 20(1):101. [PubMed: 21572976]
- Felsher, Dean W.; Zetterberg, Anders; Zhu, Jiyue; Tlsty, Thea; Michael Bishop, J. Over-expression of myc causes p53-dependent g2 arrest of normal fibroblasts. *Proceedings of the National Academy of Sciences*. 2000; 97(19):10544–10548.
- Friedman, JH. Fast sparse regression and classification. 2008.
- Garcia RI, Ibrahim JG, Zhu H. Variable selection for regression models with missing data. *Statistica Sinica*. 2010; 20(1):149. [PubMed: 20336190]
- Hilbe, JM. Negative binomial regression. Cambridge Univ Pr; 2011.
- Ibrahim, Joseph G. Incomplete data in generalized linear models. *Journal of the American Statistical Association*. 1990; 85(411):765–769.
- Ji, Hongkai; Jiang, Hui; Ma, Wenxiu; Johnson, David S.; Myers, Richard M.; Wong, Wing H. An integrated software system for analyzing chip-chip and chip-seq data. *Nat Biotechnol*. 2008; 26(11):1293–1300. [PubMed: 18978777]
- Khalili A, Chen J. Variable selection in finite mixture of regression models. *Journal of the American Statistical Association*. 2007; 102(479):1025–1038.
- Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, Zhang MQ, Lobanenko VV, Ren B. Analysis of the vertebrate insulator protein ctf binding sites in the human genome. *Cell*. 2007; 128(6):1231. [PubMed: 17382889]
- Kolasinska-Zwiercz, Paulina; Down, Thomas; Latorre, Isabel; Liu, Tao; Shirley Liu, X.; Ahringer, Julie. Differential chromatin marking of introns and expressed exons by h3k36me3. *Nature genetics*. 2009; 41(3):376–381. [PubMed: 19182803]
- Kuan PF, Chung D, Pan G, Thomson JA, Stewart R, Kele S. A statistical framework for the analysis of chip-seq data. *Journal of the American Statistical Association*. 2011; 106(495):891–903.
- Li, Bing; Gogol, Madelaine; Carey, Mike; Lee, Daeyoup; Seidel, Chris; Workman, Jerry L. Combined action of phd and chromo domains directs the rpd3s hdac to transcribed chromatin. *Science*. 2007; 316(5827):1050–1054. [PubMed: 17510366]
- Li, Feng; Mao, Guogen; Tong, Dan; Huang, Jian; Gu, Liya; Yang, Wei; Li, Guo-Min. The histone mark h3k36me3 regulates human dna mismatch repair through its interaction with muts α . *Cell*. 2013; 153(3):590–600. [PubMed: 23622243]
- Liu, Chuan-ju; Prazak, Lisa; Fajardo, Marc; Yu, Shuang; Tyagi, Neetu; Di Cesare, Paul E. Leukemia/lymphoma-related factor, a pox domain-containing transcriptional repressor, interacts with histone deacetylase-1 and inhibits cartilage oligomeric matrix protein gene expression and chondrogenesis. *Journal of Biological Chemistry*. 2004; 279(45):47081–47091. [PubMed: 15337766]
- Lozzio, Carmen B.; Lozzio, Bismarck B. Human chronic myelogenous leukemia cell-line with positive philadelphia chromosome. *Blood*. 1975; 45(3):321–334. [PubMed: 163658]

- Lystig TC, Hughes JP. Exact computation of the observed information matrix for hidden markov models. *Journal of Computational and Graphical Statistics*. 2002; 11(3):678–689.
- McLachlan GJ. On the em algorithm for overdispersed count data. *Stat Methods Med Res*. Mar; 1997 6(1):76–98. [PubMed: 9185291]
- Meng, Xiao-Li; Rubin, Donald B. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*. 1993; 80(2):267–278.
- Qin ZS, Yu J, Shen J, Maher CA, Hu M, Kalyana-Sundaram S, Yu J, Chinnaiyan AM. Hpeak: an HMM-based algorithm for defining read-enriched regions in ChIP-seq data. *BMC Bioinformatics*. 2010; 11:369. [PubMed: 20598134]
- Rashid NU, Giresi PG, Ibrahim JG, Sun W, Lieb JD. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biology*. 2011; 12(7):R67. [PubMed: 21787385]
- Roberts, Charles WM.; Orkin, Stuart H. The swi/snf complex chromatin and cancer. *Nature Reviews Cancer*. 2004; 4(2):133–142.
- Rozowsky, Joel; Euskirchen, Ghia; Auerbach, Raymond K.; Zhang, Zhengdong D.; Gibson, Theodore; Bjornson, Robert; Carriero, Nicholas; Snyder, Michael; Gerstein, Mark B. Peakseq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotech*. 2009; 27(1):66–75.
- Schwarz, Gideon. Estimating the dimension of a model. *The Annals of Statistics*. 1978; 6(2):461–464.
- Spyrou, Christiana; Stark, Rory; Lynch, Andy G.; Tavaré, Simon. Bayespeak: Bayesian analysis of ChIP-seq data. *BMC Bioinformatics*. 2009; 10:299. [PubMed: 19772557]
- Sun W, Ibrahim JG, Zou F. Genomewide multiple-loci mapping in experimental crosses by iterative adaptive penalized regression. *Genetics*. 2010; 185(1):349. [PubMed: 20157003]
- Suvà, Mario L.; Riggi, Nicolo; Bernstein, Bradley E. Epigenetic reprogramming in cancer. *Science*. 2013; 339(6127):1567–1570. [PubMed: 23539597]
- Thurman, Robert E.; Rynes, Eric; Humbert, Richard; Vierstra, Jeff; Maurano, Matthew T.; Haugen, Eric; Sheffield, Nathan C.; Stergachis, Andrew B.; Wang, Hao; Vernot, Benjamin, et al. The accessible chromatin landscape of the human genome. *Nature*. 2012; 489(7414):75–82. [PubMed: 22955617]
- Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1996:267–288.
- Wang, Gang G.; Cai, Ling; Pasillas, Martina P.; Kamps, Mark P. Nup98–nsd1 links h3k36 methylation to hox-a gene activation and leukaemogenesis. *Nature cell biology*. 2007; 9(7):804–812.
- Xu H, Wei CL, Lin F, Sung WK. An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. *Bioinformatics*. 2008; 24(20):2344–9. [PubMed: 18667444]
- Zeger SL, Qaqish B. Markov regression models for time series: a quasi-likelihood approach. *Biometrics*. 1988; 44(4):1019–31. [PubMed: 3148334]
- Zeger SL. A regression model for time series of counts. *Biometrika*. 1988; 75(4):621.
- Zhang CH. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*. 2010; 38(2):894–942.
- Zhang, Yong; Liu, Tao; Meyer, Clifford A.; Eeckhoute, Jérôme; Johnson, David S.; Bernstein, Bradley E.; Nusbaum, Chad; Myers, Richard M.; Brown, Myles; Li, Wei, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. 2008a; 9(9):R137. [PubMed: 18798982]
- Zhang ZD, Rozowsky J, Snyder M, Chang J, Gerstein M. Modeling ChIP sequencing in silico with applications. *PLoS Comput Biol*. 2008b; 4(8):e1000158. [PubMed: 18725927]
- Zou H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*. 2006; 101(476):1418–1429.

A. APPENDIX

A.1 Regularity Conditions for Corollary 1

Define the standard HMM as a special case of the HMM with covariates where covariates are not utilized, $l_n(\Psi_0|\mathbf{X}, y)$ is the likelihood of the standard HMM, and Ψ_0 is the true value of Ψ . In the standard HMM, only an intercept is used to model the mean of each state distribution. Let $\hat{\Psi}$ be the MLE of Ψ and assume that the Fisher Information matrix $I(\Psi_0)$ exists and is non-singular. Then, assuming the regularity conditions A1–A6 from Bickel et al. [1998], we have the following:

1. $\hat{\Psi} \rightarrow \Psi_0$ almost surely as $n \rightarrow \infty$.
2. $n^{-\frac{1}{2}}l'_n(\Psi_0|\mathbf{X}, y) \rightarrow N(0, I(\Psi_0))$ in distribution as $n \rightarrow \infty$
3. $\frac{1}{n}l''_n(\hat{\Psi}|\mathbf{X}, y) \rightarrow -I(\Psi_0)$ in probability as $n \rightarrow \infty$
4. $n^{\frac{1}{2}}(\hat{\Psi} - \Psi_0) \rightarrow N(0, I(\Psi_0^{-1}))$ in distribution as $n \rightarrow \infty$

For the HMM with covariates case, $(Y_1, Z_1), \dots, (Y_n, Z_n)$ is conditioned on $X_1, \dots, X_n$ such that $Y_1|X_1, Z_1, \dots, Y_n|X_n, Z_n$ are conditionally independent and that $(Y_1, Z_1)|X_1, \dots, (Y_n, Z_n)|X_n$ is stationary conditional on X_i . We assume that our HMM with covariates is identifiable, such that for any set of parameters Ψ and Ψ^* , $l_n(\Psi^*|\mathbf{X}, y) = l_n(\Psi|\mathbf{X}, y)$ if and only if $\Psi^* = \Psi$ and $K^* = K$ up to a permutation of the states. The above results demonstrate that the HMM with covariates likelihood have similar asymptotic properties to the typical *iid* likelihood. Given results (1)–(4) and because penalty conditions P0–P2 are similar to those from Khalili and Chen [2007], Corollary 1 naturally follows from Khalili and Chen [2007] Theorems 1 and 2. To avoid duplicating the proof from Khalili and Chen [2007], we describe the proof of Corollary 1 as follows. The proof of Corollary 1, part (a), follows from Khalili and Chen [2007] Theorem 1 with results (2) and (3) above. The proof of Corollary 1, part (b), follows from Khalili and Chen [2007] Theorem 2 part a and b.1 with regularity condition A3 and result (4). Proof of part (c) of Corollary 1 follows from Khalili and Chen [2007] b.2 with regularity condition results (2) and (3).

A.2 AR-HMM Forward, Backward, and related probabilities

For any $k = 1, \dots, K$, define

$$\begin{aligned} f_{1k} &= p(Z_1=k, y_1|\mathbf{X}, \Psi_A^{(s)}) = \pi_k f_k(y_1|X_{1k}, \Psi_A^{(s)}), \\ f_{ik} &= p(Z_i=k, y_1, \dots, y_i|\mathbf{X}, \Psi_A^{(s)}) = \sum_{j=1}^K \alpha_{i-1,j} \gamma_{jk} f_{jk}(y_i|X_{ik}, r_{i-1,j}^{(s)}, \Psi_A^{(s)}), \text{ for } i > 1, \\ b_{ik} &= p(y_{i+1}, \dots, y_n|Z_i=k, \mathbf{X}, \Psi_A^{(s)}) = \sum_{j=1}^K \gamma_{kj} f_{kj}(y_{i+1}|X_{i+1,j}, r_{i,k}^{(s)}, \Psi_A^{(s)}) b_{i+1,j}, \text{ for } i < n, \end{aligned}$$

and $b_{nk} = 1$. Then we have

$$p(Z_i=k|y, \mathbf{X}, \Psi_A^{(s)}) = \frac{f_{ik} b_{ik}}{\sum_{l=1}^K f_{nl}}, \quad (\text{A.1})$$

$$p(Z_{i-1}=j, Z_i=k|y, \mathbf{X}, \Psi_A^{(s)}) = \frac{f_{i-1,j} \gamma_{jk}^{(s)} f_{jk}(y_i | X_{ik}, r_{i-1,j}^{(s)}, \Psi_A^{(s)}) b_{ik}}{\sum_{l=1}^K f_{nl}}. \quad (\text{A.2})$$

Figure 1.

Autocorrelation of window (250 bp windows) read counts from background (A, C) and enriched regions (B, D) of CTCF (CCCTC-binding factor) ChIP-seq (A, B) and H3K36me3 (Trimethylation of Lys36 in histone H3) ChIP-seq (C, D) dataset measured in Chr22 of a human cell line (GM12878). Window read counts were log transformed. Likely enriched regions were determined by fitting a two-component Negative Binomial Finite Mixture Model and regions classified to be enriched at an FDR threshold of 0.05.

Figure 2.

Comparison of classification performance (at window level) of the AR-HMM, HMM, and FMR for $\nu_1 = 0.2$ and $\nu_2 = 0.2$ (first column, low auto-correlation in enriched regions) and $\nu_1 = 0.2$ and $\nu_2 = 0.8$ (second column, high auto-correlation in enriched regions). The first row are the results for CTCF-style ChIP-seq data with short enriched regions, and the second row are the results for H3K36me3 histone modification-style data with longer regions of enrichment.

Figure 3.

Classification performance comparison of the FMR, HMM, and AR-HMM models in GM12878 CTCF ChIP-seq. A) Number of significantly enriched regions (which are generated by collapsing adjacent significant windows) called by each method across FDR thresholds. B) Average length of significant regions across FDR thresholds. C) Classification performance relative to FDR threshold, where regions overlapping a CTCF binding motif are classified as “correct”. D) Box plots of maximum window read counts from significant regions called uniquely by each method (the first three box plots) and box plot of maximum window read count from background called by all three methods (the last box plot). E–F) Examples of enriched regions called by each method at FDR level 0.05.

Figure 4. Classification performance comparison of the FMR, HMM, and AR-HMM models in GM12878 H3K36me3 ChIP-seq data. A) The total length of significantly enriched regions (which are generated by collapsing adjacent significant windows) that overlap a gene body across FDR thresholds. B) Average lengths of significant regions across FDR thresholds. C) Number of genes overlapped with significant regions across FDR thresholds. D) Median proportion of gene bodies covered by significantly enriched regions. E–F) Examples of regions called as enriched by each method.

Figure 5.

The variable selection results for background state and enriched state of H3K36me3. Each variable represent the binding signals of a transcription factor (TF).

Table 1

Two-state Poisson HMM variable selection simulation setup for regression coefficients corresponding to states 1 and 2 (background vs. enriched).

	β_1	β_2	β_3	β_4
State 1	2	0	0	0
State 2	0	2	2	2

Table 2

Variable selection performance in simulated two-state Poisson HMM with covariates over a range of conditions. p : simulated number of covariates per state; TD: average number of “True Discoveries” (equal to 4); FD: average number of “False Discoveries”; the remaining columns are the mean estimates of the true non-zero coefficients in the model from Table 1.

Type	p	n	Pen	TD(4)	FD(0)	β_{11}	β_{22}	β_{23}	β_{24}	γ_{11}	γ_{12}	γ_{21}	γ_{22}
CTCF	5	200	lasso	3.99	1.81	1.79	1.88	1.93	1.91	0.894	0.106	0.891	0.109
			log	3.91	0.33	1.93	1.93	1.96	1.93	0.896	0.104	0.904	0.0956
		10000	scad	4	0.43	2	2.01	2	1.98	0.899	0.101	0.903	0.0968
			lasso	4	0.99	1.96	1.98	1.97	1.98	0.901	0.0994	0.901	0.0994
	100	200	log	4	0.14	1.99	2	2	2	0.899	0.101	0.9	0.0997
			scad	4	0.18	2	2	2	2	0.9	0.0998	0.901	0.0989
		10000	lasso	2.94	14.00	1.46	0.705	0.585	0.572	0.911	0.0887	0.913	0.0873
			log	3.01	7.53	1.9	1.13	0.937	0.836	0.909	0.0905	0.917	0.0829
	5	200	scad	3.1	8.44	1.92	1.55	1.24	1.05	0.89	0.11	0.9	0.0998
			lasso	4	1.78	1.92	1.92	1.93	1.92	0.903	0.0973	0.902	0.0977
Histone	5	200	log	4	0.43	2	2	2	2	0.9	0.0999	0.9	0.0997
			scad	4	0.18	2	2	2	2	0.899	0.101	0.901	0.0988
		10000	lasso	4	1.55	1.71	1.95	1.96	1.94	0.897	0.103	0.117	0.883
			log	3.97	0.15	1.92	1.99	1.99	1.99	0.894	0.106	0.117	0.883
	100	200	scad	4	0.14	1.97	1.99	1.99	2.01	0.9	0.0999	0.113	0.887
			lasso	4	0.56	1.95	1.99	1.99	1.99	0.9	0.0998	0.0999	0.9
		10000	log	4	0.11	2	2	2	2	0.9	0.1	0.101	0.899
			scad	4	0.03	2	2	2	2	0.9	0.0998	0.101	0.899
	5	200	lasso	4	7.05	1.38	1.78	1.77	1.78	0.894	0.106	0.106	0.894
			log	3.98	2.98	1.95	1.98	1.99	1.95	0.9	0.0995	0.111	0.889
		10000	scad	4	4.94	1.99	2.01	1.99	2	0.903	0.0967	0.106	0.894
			lasso	4	0.95	1.91	1.97	1.97	1.97	0.9	0.1	0.1	0.9
Histone	5	200	log	4	0.45	2	2	2	2	0.9	0.0998	0.101	0.899
			scad	4	0.04	2.01	2	2	2	0.9	0.1	0.1	0.9
		10000	lasso	4	0.95	1.91	1.97	1.97	1.97	0.9	0.1	0.1	0.9
			log	4	0.45	2	2	2	2	0.9	0.0998	0.101	0.899
	100	200	lasso	4	7.05	1.38	1.78	1.77	1.78	0.894	0.106	0.106	0.894
			log	3.98	2.98	1.95	1.98	1.99	1.95	0.9	0.0995	0.111	0.889
		10000	scad	4	4.94	1.99	2.01	1.99	2	0.903	0.0967	0.106	0.894
			lasso	4	0.95	1.91	1.97	1.97	1.97	0.9	0.1	0.1	0.9
	5	200	log	4	0.45	2	2	2	2	0.9	0.0998	0.101	0.899
			scad	4	0.04	2.01	2	2	2	0.9	0.1	0.1	0.9

Mean parameter estimates (of 1000 simulations) for the AR-HMM, HMM, and FMR models applied to simulated two-state Poisson AR-HMM ChIP-seq data of CTCF binding sites (CTCF) and H3K36me3 histone modifications (Histone) over various conditions. The number of windows $n = 10,000$.

Table 3

Type	$v_{1,r}$	$v_{2,r}$	Model	β_{10}	β_{11}	β_{21}	v_1	β_{02}	β_{12}	β_{22}	v_2	γ_{11}	γ_{12}	γ_{21}	γ_{22}
CTCF	0.2	0.2	ARH	0	1	0.999	0.2	1.5	2	2	0.2	0.9	0.1	0.9	0.1
	0.2	0.2	HMM	-0.017	0.999	0.999		1.484	1.998	1.999		0.9	0.1	0.9	0.1
	0.2	0.2	FMR	-0.017	0.999	0.999		1.484	1.998	1.999					
	0.2	0.8	ARH	-0.001	1	1.001	0.2	1.5	2	2	0.8	0.9	0.1	0.9	0.1
	0.2	0.8	HMM	0.011	0.985	0.985		1.573	1.939	1.942		0.907	0.093	0.903	0.097
	0.2	0.8	FMR	0.011	0.985	0.985		1.573	1.939	1.942					
Histone	0.2	0.2	ARH	0.001	0.999	0.999	0.199	0.5	1.999	2	0.2	0.9	0.1	0.101	0.899
	0.2	0.2	HMM	-0.026	1.006	1.006		0.497	1.995	1.996		0.9	0.1	0.101	0.899
	0.2	0.2	FMR	-0.021	1.004	1.002		0.495	1.996	1.997					
	0.2	0.8	ARH	-0.002	1.001	1.002	0.199	0.5	2	2	0.8	0.9	0.1	0.1	0.9
	0.2	0.8	HMM	-0.052	1.102	1.104		0.629	1.921	1.922		0.902	0.098	0.129	0.871
	0.2	0.8	FMR	0.021	1.051	1.053		0.655	1.904	1.905					

Table 4

Variable selection performance and estimation accuracy for true non-zero coefficients in simulated two-state Poisson AR-HMM with covariates data over a range of conditions and $\nu_1 = \nu_2 = 0.4$

Type	p	n	Pen	TD(4)	FD(0)	β_{11}	ν_1	β_{22}	β_{23}	β_{24}	ν_2	γ_{11}	γ_{21}	γ_{12}	γ_{22}
CTCF	5	200	lasso	3.97	2.06	1.79	0.367	1.84	1.88	1.81	0.189	0.909	0.0909	0.92	0.0799
			log	3.86	0.43	1.96	0.367	1.89	1.85	1.91	0.227	0.902	0.0982	0.927	0.0729
			scad	3.99	0.58	1.97	0.386	1.97	2.06	1.95	0.11	0.902	0.0979	0.911	0.0894
		10000	lasso	4	1.18	1.96	0.396	1.98	1.98	1.98	0.388	0.901	0.0994	0.898	0.102
			log	4	0.05	2	0.393	2	2	2	0.386	0.9	0.1	0.901	0.0989
			scad	4	0.03	2	0.391	2	2	2	0.388	0.9	0.0997	0.899	0.101
	100	200	lasso	2.59	17.11	1.194	0.352	0.468	0.47	0.418	0.218	0.865	0.135	0.877	0.123
			log	2.69	9.7	1.657	0.355	0.936	0.832	0.917	0.034	0.864	0.136	0.876	0.124
			scad	3.32	12.11	1.315	0.342	1.519	1.525	1.315	-0.049	0.85	0.15	0.861	0.139
		10000	lasso	4	2.49	1.931	0.403	1.929	1.928	1.932	0.402	0.902	0.098	0.9	0.1
			log	4	0	1.976	0.402	1.996	1.996	1.995	0.401	0.9	0.1	0.9	0.1
			scad	4	0.23	1.994	0.401	2	2	2.003	0.401	0.9	0.1	0.9	0.1
Histone	5	200	lasso	4	1.8	1.72	0.387	1.97	1.96	1.94	0.395	0.893	0.107	0.113	0.887
			log	3.97	0.31	1.93	0.379	2	2	1.99	0.39	0.898	0.102	0.108	0.892
			scad	4	0.2	1.95	0.384	2.01	1.99	1.98	0.383	0.901	0.0987	0.103	0.897
		10000	lasso	4	0.72	1.96	0.401	1.99	1.99	1.99	0.401	0.9	0.0996	0.101	0.899
			log	4	0.1	2	0.401	2	2	2	0.401	0.9	0.1	0.101	0.899
			scad	4	0.03	2	0.4	2	2	2	0.4	0.901	0.0995	0.1	0.9
	100	200	lasso	4	19.07	1.33	0.367	1.812	1.829	1.799	0.386	0.896	0.104	0.114	0.886
			log	4	5.57	1.856	0.382	1.978	1.957	1.973	0.389	0.908	0.092	0.107	0.893
			scad	4	16.48	1.847	0.344	1.991	1.98	1.986	0.337	0.897	0.103	0.117	0.883
		10000	lasso	4	1.93	1.909	0.407	1.974	1.973	1.974	0.404	0.901	0.099	0.101	0.899
			log	4	0.03	1.986	0.4	1.985	1.985	1.985	0.401	0.901	0.099	0.101	0.899
			scad	4	0.2	1.999	0.397	2.001	2.001	2.001	0.402	0.9	0.1	0.1	0.9

Table 5

GM12878 CTCF and H3K36me3 ChIP-seq Chr22 two-state Negative Binomial HMM and AR-HMM real data variable selection results. $\beta_{0,k}$ is the intercept in state k , $\beta_{1,k}$ corresponds to the GC content main effect, $\beta_{2,k}$ corresponds to the mappability main effect, $\beta_{3,k}$ corresponds to the input control main effect. Interaction terms are denoted with combination of indices, for example $\beta_{12,k}$ corresponds to GC content-mappability interaction term.

Data	CTCF	CTCF	Histone	Histone
Model	HMM	ARH	HMM	ARH
BIC	386889	384512	545036	532506
$\beta_{0,1}$	-2.273	-2.318	0.270	0.179
$\beta_{1,1}$	1.611	1.613	1.900	2.204
$\beta_{2,1}$	1.277	1.387	0.241	0.325
$\beta_{3,1}$			0.072	0.0617
$\beta_{12,1}$			-2.707	-3.039
$\beta_{13,1}$				
$\beta_{23,1}$				
$\beta_{123,1}$	0.628	0.57	0.256	0.245
v_1		0.283		0.161
$\beta_{0,2}$	-2.557	-3.171	1.315	1.071
$\beta_{1,2}$	5.91	6.987		0.521
$\beta_{2,2}$	2.123	2.201	0.545	0.734
$\beta_{3,2}$	0.597	0.557	0.235	0.193
$\beta_{12,2}$			0.377	
$\beta_{13,2}$	0.002			
$\beta_{23,2}$				
$\beta_{123,2}$			-0.297	-0.258
v_2		0.749		0.583

Table 6

Proportion of significant regions from each method (columns) that overlap with peaks from other methods (FDR=0.05). Cells corresponding to the same method are those that unique only to that method. For example, 92% and 88% of the 1180 significant FMR regions (Column 1, Rows 2 and 3) overlap with the HMM and AR-HMM, respectively.

	FMR	HMM	ARHMM
FMR	1	0.95	0.97
HMM	0.92	1	0.96
ARH	0.88	0.93	1
Regions	1180	1039	962

Table 7

Chr22 GM12878 CTCF and H3K36me3 ChIP-seq penalized model estimates - transition probabilities and dispersion parameters

Data	Type	η_{11}	η_{12}	η_{21}	η_{22}	ϕ_1	ϕ_2
CTCF	ARH	0.988	0.012	0.183	0.817	1.095	0.637
CTCF	HMM	0.989	0.011	0.255	0.745	1.061	0.563
CTCF	FMR					1.136	0.105
H3K36me3	ARH	0.997	0.003	0.006	0.994	5.567	6.748
H3K36me3	HMM	0.995	0.005	0.012	0.988	4.752	4.874
H3K36me3	FMR					4.656	3.779