A Bayesian Graphical Model for ChIP-Seq Data on Histone Modifications

Riten MITRA[1], Peter MÜLLER[2*], Shoudan LIANG[3], Lu YUE[4], Yuan JI[5*]

[1] ICES, University of Texas at Austin

[2] Department of Mathematics, University of Texas at Austin

[3] Department of Bioinformatics and Computational Biology,

The University of Texas, MD Anderson Cancer Center

[4] Department of Leukemia,

The University of Texas, MD Anderson Cancer Center

[5] Center for Clinical and Research Informatics,

NorthShore University HealthSystem

August 10, 2012

*To whom correspondence should be addressed:pmueller@math.utexas.edu, koaeraser@gmail.com

Abstract

Histone modifications (HMs) are an important post-translational feature. Different types of HMs are believed to co-exist and co-regulate biological processes such as gene expression, and therefore are intrinsically dependent on each other. We develop inference for this complex biological network of HMs based on a graphical model using ChIP-Seq data. A critical computational hurdle in the inference for the proposed graphical model is the evaluation of a normalization constant in an autologistic model that builds on the graphical model. We tackle the problem by Monte Carlo evaluation of ratios of normalization constants. We carry out a set of simulations to validate the proposed approach and to compare it with a standard approach using Bayesian networks. We report inference on HM dependence in a case study with ChIP-Seq data from a next-generation sequencing experiment. An important feature of our approach is that we can report coherent probabilities and estimates related to any event or parameter of interest, including honest uncertainties. Posterior inference is obtained from a joint probability model on latent indicators for the recorded HMs. We illustrate this in the motivating case study. An R package including an implementation of posterior simulation in C is available.

KEYWORDS: Auto logistic regression; Epigenetics; Histone modifications; Markov chain Monte Carlo; Markov random fields; Network model; Pathway dependence.

1 INTRODUCTION

Genes are expressed in eukaryotes through complex processes that involve multiple layers of control. In addition to transcription factors that activate or repress their target genes by binding to DNA response elements in the promoters of the genes, the eukaryotic cells have additional layers of control of gene expression by chemical modifications to the histones. Histones are small proteins; they form a core around which DNA is wrapped, forming

nucleosomes. The combination of DNA and nucleosomes is called chromatin (Figure 1). Nucleosomes are the basic *in-vivo* structural unit of DNA, similar to thread wrapped around a stool. The thread is 147 base pairs of DNA and the stool consist of an octamer of the four core histones (two sets of H2A, H2B, H3, and H4). While the role of H2A and H2B are primarily structural, histones H3 and H4 play key roles in integrating a variety of signals that regulate gene transcription. In particular, these two histone proteins have "tails" — strings of amino acids that protrude outside of the basic nucleosomal structure and make contact with DNA. Specific post-translational modifications of the amino acids in these histone tails (*e.g.* methylation, acetylation, phosphorylation, sumoylation, and ubiquination) interact with other proteins to either relax the chromatin and promote transcription, or to condense the chromatin into a closed form, which excludes transcription factors and result in gene silencing. Majority of these *histone modifications* (HMs) are epigenetic, *i.e.* that they are faithfully preserved in DNA replications. Thus daughter cells inherit the same patterns of HMs that occur in the parent cell. A comprehensive list of histone modifications appears at `http://bioinfo.hrbmu.edu.cn/hhmd` (Zhang et al., 2010).

The correlation of HMs with translational activity and occurrence of promoters has been been well documented in Bernstein et al. (2002), Kim et al. (2005) and Roh et al. (2005). For example, histone methylations influence gene activation and repression, and histone acetylations are associated with a variety of functions such as gene activation, nucleosome assembly, higher-order chromatin packing and interactions with non-histone proteins (Grant and Berger, 1999). HMs also play an influential role in DNA damage repair and chromosomal segregation. See Bergink et al. (2006) and Andersson et al. (2009) for details.

Biochemically, different enzymes can facilitate different modifications to histones in parallel, thus making it possible for the co-existence of multiple HMs. It is hypothesized that the combination of different HMs dictates the status of a gene being permissible to transcription (active) or not (repressed) (Strahl and Allis, 2000). The hypothesis is known as

the *histone code*. Understanding the histone code is among the most exciting challenges facing scientists in genomics research today.

Motivated by the histone code hypothesis, researchers have attempted to demarcate functional domains over the genome by signatures of histone patterns (Liu et al., 2005; Pokholok et al., 2005; Heintzman et al., 2009) and reported some interesting findings. However, a full description of the histone code remains elusive until now. Attempts to decode the combination of HMs using quantitative methods have been primarily based on descriptive statistics, and independent and HM-specific hypothesis tests. For example, Wang et al. (2008) tested the enrichment of 39 individual HMs across the genome and identified a set of 17 HMs with largest p-values based on a Poisson assumption for the HM counts from a ChIP-Seq experiment. They used a Bonferroni correction to adjust for the multiple comparisons. They also computed pairwise correlations between any two HMs and conjectured that the set of these 17 HMs serves as a backbone of the dynamics of the chromatin structure genome wide and plays an important role in gene regulation.

Building on these results, we propose to apply graphical models to investigate the complex dependence relationship of multiple HMs. To our knowledge, this is the first attempt to characterize the interaction of the HMs based on probabilistic graphical models. These models will allow us to identify the type and the strength of interaction of co-existing HMs. More importantly, the models provide a full probabilistic description of any subgraph involving subsets of HMs. For example, we present in Section 7.1 a highly connected graph involving the 17 backbone HMs. Our results confirm the findings in Wang et al. (2008) and elaborate their results with a full descriptions of all interactions. In addition, applying the same graphical model to the full list of 39 HMs in Wang et al. we reveal new insights in the entire HM-interaction map, thus making a step forward towards filling in part of the puzzle of the histone code. For the 17 backbone HMs, we shall focus only on dependence across HM's, assuming independence across loci. In the full data analysis we will restrict inference

to coding regions only, thus reducing possible spatial correlations.

In the next section we give a brief overview of a ChIP-Seq experiment that was carried out to record the HMs, and we discuss the statistical challenges posed by such data. In Section 3 we introduce the concept of Markov random fields (MRF) as a tool for describing dependence. Section 4 elaborates on the proposed probability model and Section 5 describes posterior inference. In Section 6 we present a simulation experiment to evaluate the performance of the proposed graphical model. In Section 7 we report results of the analysis of the ChIP-Seq data. Finally, we conclude with a discussion in Section 8.

2 ChIP-SEQ DATA

We analyze data from a ChIP-seq experiment for CD4+ T lymphocytes (Barski et al., 2007; Wang et al., 2008). ChIP-Seq is a new technique that combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing (Seq) to identify genome-wide binding patterns of DNA-associated proteins, such as HMs. In such an experiment, HM-specific antibodies are used to preferentially extract the histones along with the DNA associated with it. DNA that cross-linked with the specific HM is then randomly broken up, by sonication, into pieces of several hundred nucleotides. The DNA pieces are used to generate short reads, which are sequenced and mapped to the genome. At the end of a ChIP-Seq experiment, the final measurement is the counts of DNA short reads that are mapped to specific loci on the genome. A loci here refers to a DNA segment of around 2,000 base pairs. The larger the DNA short reads j count, the higher the amount of the HM. By using 39 different antibodies and carrying out a ChIP-Seq experiment for each HM, Barski et al. (2007) and Wang et al. (2008) report HMs counts for $m = 39$ types of HMs including 18 acetylations, 20 methylations, and one special histone modification H2A.Z.

We keep all genomic locations with at least one enriched HM and drop all the windows

where none of the 39 HMs are enriched. For the purpose of this screening, we use the peak-calling program SICER (Zang et al. 2009) to decide enrichment. SICER parameters were set to `W_SIZE=200, GAP_SIZE=600, EVALUE=1000, FRAG_SIZE=150`. Also any adjacent windows with unchanged SICER calls for the 39 HM counts were merged to create larger regions.

The data is arranged in an $m \times n$ matrix $\boldsymbol{Y} = [y_{it}]$ of counts. Each row, $i = 1, \ldots, m$ represents a type of HM, and each column, $t = 1, \ldots, n$, represents a genomic location. That is, the count y_{it} reports the number of HMs of type i at location t. In Sections 7.1 and 7.2, we analyze two datasets with $m = 17$ or 39, and $n = 50,000$ or $33,681$, respectively.

These HM counts are results of upstream data preprocessing and normalization, which are themselves important research topics (Kuan et al., 2011; Rozowsky et al., 2009), but are beyond the scope of this discussion. In addition, our approach does not infer enrichment of histone counts over the genome, which is another concurrent topic in ChIP-Seq data analysis (e.g., Kuan et al, 2011). Our goal is to study the dependence of HMs as an inference problem downstreams of data normalization and peak detection.

3 MODELING DEPENDENCE

3.1 Network Models

Traditional inference for dependence structure is often implemented through Bayesian networks based on Gaussian graphical models. However, there are important limitations that make it impractical to use this approach for the desired analysis of the HM data. First, the estimation of the reported network is based on a greedy heuristic search. In contrast, the proposed approach allows for a full model-based and probabilistic inference. For example, under the posterior inference we can provide error summaries such as false discovery rates

for each estimated graph; we can also report full posterior inference on any subgraphs of interest. Second, the multivariate normal sampling model implied in a Bayesian network is not appropriate for the application to HM data. To be a useful tool for decoding the histone code, the sampling model needs to include biologically meaningful indicators for presence versus absence of an HM.

We propose an extension of traditional Bayesian network models to address these limitations. In particular, we propose a model and inference approach that relies on data reduction to a binary latent biological signal e_{it}. This latent binary variable codes for presence versus absence of HMs. The reduction to a binary underlying signal is similar to the probability of expression (POE) model proposed in Parmigiani et al. (2002) for microarray data.

Our proposed model for HM's is based on a Markov random field (MRF) graph. We define an MRF graph as a pair $G = (V, E)$, where V is a set of vertices and E is a set of undirected edges. The vertices correspond to the variables, in our case HMs. The edges are a subset of $\{\{v_1, v_2\},\ v_1 \neq v_2 \in V\}$. The absence of an edge $\{v_i, v_j\}$ indicates conditional independence of the corresponding variables. The Hammersley Clifford Theorem assures us that any given conditional independence structure can be represented by an MRF. See Besag (1974). The MRF consists of undirected edges. Note that the proposed approach does not allow inference on directed graphs. For a review of some interesting alternative approaches that allow inference about directed graphs see, for example, Kalisch and Bühlmann (2007).

3.2 Prior on Random Graphs

We start the model construction for the HM data with a prior $p(G)$ on the dependence structure represented by an MRF graph $G = (V, E)$. We propose two types of prior constructions depending on the size of the graph. For smaller graphs, such as the network with the selected $m_1 = 17$ HMs described in Section 2, we define a prior $p(G)$ as a uniform distribution over

all possible subgraphs with vertex set V including all 17 HMs. The proposed uniform prior is only practical for a moderately small number of vertices, say for m_1 close to 15. For larger graphs we find in simulation studies (not shown) that posterior inference can no longer reliably recover the simulation truth without bringing prior information to bear. One exception is highly sparse graphs. More importantly, when sufficient prior information based on biological expertise or related data is available, we can construct informative priors and consider inference for a larger number of vertices (Telesca et al., 2010). For example, let $G_0 = (V, E_0)$ denote an informed prior guess of the dependence structure. For applications with genomic data this is often available as a consensus pathway for a non-pathological state. Then $p(G)$ could be based on the number of changes relative to G_0, $d(G, G_0) = |E \cap E_0^c| + |E^c \cap E_0|$, using for example, a geometric distribution in d. That is, we assume

$$p(G) \propto \rho^{d(G,G_0)}. \tag{3.1}$$

We shall refer to ρ as the concentration parameter in later discussions. The prior $p(G)$ implies that we place less weight on graphs which are more distant from G_0. The weights reduce exponentially with increasing values of d. Noting that $p(G)/p(G_0) = \rho^{d(G,G_0)}$ is interpreted as prior odds, one can easily calibrate ρ to reflect the prior confidence on G_0. For example, if one believes that the true graph should not differ from G_0 by more than 20 edges, placing $\rho = 0.90$ would result in a prior odds of $0.90^{20} \approx 0.12$ for a graph G with 20 flipped edges relative to G_0. Lastly, model (3.1) can be easily modified to incorporate weights on different edges, as a reflection of the reliability of the prior knowledge. For example, edges that are experimentally validated could have larger weights in computing the distance function $d()$.

4 MODELING HMs

4.1 A Prior Model for HM Indicators e_{it}

The proposed model for HM data is constructed as a hierarchical model, starting with the prior $p(G)$. The remaining layers of the hierarchical model are introduced one at a time in this section. For reference we state the overall model structure

$$p(G)\, p(\boldsymbol{\beta} \mid G)\, p(\boldsymbol{e} \mid \boldsymbol{\beta}, G)\, p(\boldsymbol{y} \mid \boldsymbol{e}). \tag{4.1}$$

The first factor is the prior on G. The next two factors define a prior on latent binary indicators e_{it} for presence of histone modifications. The last layer of the hierarchical model is a sampling model for the observed counts conditional on the latent indicators.

We next discuss the specification of $p(\boldsymbol{e} \mid \boldsymbol{\beta}, G)$. The problem of constructing a joint probability distribution that honors the dependence structure of the graphical model is greatly simplified by reducing the count data y_{it} to an underlying binary latent variable. Let $e_{it} \in \{0, 1\}$ denote an indicator for the presence of the i-th HM in the t-th genomic location. There are two main motivations for introducing the latent indicators e_{it}. First, the presence of an HM is a biologically meaningful variable. Second, and at least equally importantly, for the binary indicator variables e_{it} it is possible to consider the family of all possible probability models. For the moment we drop the sample index $_t$ in e_{it} to simplify notation while we discuss the joint probability model for the vector of indicators. Besag (1974) gives a convenient parameterization of all possible joint probability models for a multivariate binary vector $\boldsymbol{e} = (e_1, \ldots, e_m) \in \{0, 1\}^m$ that obey a given conditional independence structure of

an MRF. Given G and a set of coefficients $\boldsymbol{\beta}$ that index the models we have

$$p(\boldsymbol{e} \mid \boldsymbol{\beta}, G) = p(\mathbf{0} \mid \boldsymbol{\beta}, G) \times \exp\left\{\sum_i \beta_i e_i + \sum_{i<j} \beta_{ij} e_i e_j + \sum_{i<j<k} \beta_{ijk} e_i e_j e_k + \ldots + \beta_{1\cdots m} e_1 \cdots e_m\right\}, \quad (4.2)$$

known as the autologistic model. If desired, additional covariates could be included in the first term. The conditional independence structure implied in G is realized by a restriction on $\boldsymbol{\beta}$.

For any set of vertices $i_1, \ldots, i_k$ that do not form a *clique* in the graph G the corresponding interaction coefficient $\beta_{i_1 \cdots i_k}$ is zero. A clique is a set of vertices of which all pairs of vertices in the set are connected, i.e., $\{i_1, i_2\} \in E$ for all i_1, i_2 in the set. Henceforth, we use $\boldsymbol{\beta}$ to denote the vector of all non-zero coefficients $\beta_{i_1 \ldots i_k}$. The inclusion of G in the conditioning subset in (4.2) highlights the dependence of the autologistic model on G. For our implementation we assume that all interactions of order 3 and higher are zero. If desired, it is straightforward to allow for some non-zero higher order interactions. But it is impractical to consider *all* possible higher order interactions.

Caragea and Kaiser (2009) and Hughes et al. (2010) proposed a centered parameterization of the autologistic model (4.2) and argued that the centered version improves mixing of the Markov chain Monte Carlo (MCMC) posterior simulation and simplifies prior specification. Simulation studies in our application confirmed their argument. Denote with ν_i the log odds for e_i that would be implied if the i-th vertex were to share no edges with other vertices, $\nu_i = \exp(\beta_i)/\{1 + \exp(\beta_i)\}$. Then redefine (4.2) as

$$p(\boldsymbol{e}_t \mid \boldsymbol{\beta}, G) = p(\mathbf{0} \mid \boldsymbol{\beta}, G) \cdot \exp\left\{\sum_i \beta_i e_{it} + \sum_{i<j} \beta_{ij} (e_{it} - \nu_i)(e_{jt} - \nu_j)\right\}. \quad (4.3)$$

The model introduces no additional parameters. The centering $\nu_j = \nu_j(\beta_j)$ is a deterministic

function of β_j only. There is a one-to-one mapping between the coefficients of the uncentered model (4.2) and the centered model (4.3). However, despite the matching notation, the interpretation of β_i and β_{ij} in the two models differs.

The discussion in Besag (1974) implicitly assumes a known conditional independence structure G and known coefficients $\boldsymbol{\beta}$. When considering $\boldsymbol{\beta}$ and G as unknown quantities with a hierarchical prior, we run into an additional complication. Evaluation of the posterior distribution, including the complete conditional posterior distribution for $\boldsymbol{\beta}$, requires the evaluation of the normalization constant $p(\mathbf{0} \mid \boldsymbol{\beta}, G)$ in (4.3), which in turn requires a sum over all possible m-dimensional binary vectors $\boldsymbol{e}_t \in \{0,1\}^m$. The problem is well known to be computationally challenging. There are no easy solutions. Later, in Section 5 we will briefly introduce an implementation of MCMC posterior simulation that avoids the evaluation of this normalization constant and an alternative importance sampling approximation.

4.2 Sampling Model for $[y_{it}]$

We complete the model construction with a sampling distribution for the observed counts y_{it}. Inspection of the empirical distributions (not shown) of the counts for individual HMs we note several characteristic features. The histograms show typical peaks at low counts. The tails of the histograms are heavier than those of Poisson distributions, indicating against the use of a traditional Poisson sampling model for the count data. In addition, for some HMs the empirical distribution includes a second mode in the tail. This leads us to propose a mixture model, with a Poisson distribution for the low counts, say $y_{it} < c_i$, and a mixture of two log normal distributions for moderate to high counts. We interpret the peak for low counts as background when the HM is not present, i.e., under $e_{it} = 0$. Let $\text{Poi}(\lambda)$ denote a Poisson distribution with mean λ and let $\text{LN}(m, s^2)$ denote a log normal distribution with

location and scale m and s. In summary

$$p(y_{it} \mid e_{it}) \propto \begin{cases} \text{Poi}(\lambda_i)\, I(y_{it} < c_i) & e_{it} = 0 \\ \pi_i \text{LN}(\mu_{1i}, \sigma_{1i}^2) + (1 - \pi_i)\text{LN}(\mu_{2i}, \sigma_{2i}^2) & e_{it} = 1 \end{cases} \tag{4.4}$$

Let δ_x denote a point mass at x. The Poisson/log-normal mixture can be further replaced by introducing a trinary indicator $z_{it} \in \{-1, 0, 1\}$ with $p(z_{it} \mid e_{it} = 0) = \delta_{-1}(z_{it})$ and $p(z_{it} \mid e_{it} = 1) = \pi_i \delta_0(z_{it}) + (1 - \pi_i)\delta_1(z_{it})$. Then

$$p(y_{it} \mid z_{it}) = \begin{cases} \text{Poi}(\lambda_i)\, I(y_{it} < c_i) & z_{it} = -1 \\ \text{LN}(\mu_{1i}, \sigma_{1i}^2) & z_{it} = 0 \\ \text{LN}(\mu_{2i}, \sigma_{2i}^2) & z_{it} = 1 \end{cases} \tag{4.5}$$

The sampling model (4.4) assumes conditional independence of $\boldsymbol{y}$ given $\boldsymbol{e}$ across both, HM's i and loci t. This simplifying assumption reflects a preference for parsimony. We verified the assumption by considering the empirical covariance matrix of residuals and found no evidence of serious violation of the assumption, with all empirical correlations $|r_{ij}| < 0.18$. We will use $\boldsymbol{\theta} = (\pi_i, \mu_{1i}, \mu_{2i}, \sigma_{1i}, \sigma_{2i}, \lambda_i, c_i, \quad i = 1, \ldots, m)$ to denote the complete parameter vector for the sampling model. Let $\text{Ga}(a, b)$ denote a gamma distribution with mean a/b, and let $\text{Unif}(A)$ denote a uniform distribution over the set A. We assume conditionally conjugate priors

$$\lambda_i \sim \text{Ga}(\alpha, \beta),\ \mu_{ki} \sim N(0, \tau),\ p(\sigma_{ki}) \propto 1/\sigma_{ki}^2,\ \pi_i \sim Beta(1, 1),\ c_i \sim \text{Unif}(\{1, 2, 3, 4, 5\})$$

for $i = 1, \ldots, m$, $k = 1, 2$ and $\tau = 10^6$.

For later reference we state the joint model (4.1):

$$p(\boldsymbol{Y}, \boldsymbol{z}, \boldsymbol{e}, \boldsymbol{\theta}, G) \propto \underbrace{p(\boldsymbol{Y} \mid \boldsymbol{z}, \boldsymbol{\theta})}_{(4.5)}\, p(\boldsymbol{z} \mid \boldsymbol{e}, \boldsymbol{\theta})\, \underbrace{p(\boldsymbol{e} \mid \boldsymbol{\beta}, G)}_{(4.3)}\, p(\boldsymbol{\theta})\, p(\boldsymbol{\beta} \mid G)\, p(G) \tag{4.6}$$

The first factor is the sampling model (4.5). The second factor defines the trinary indicators z_{it} given e_{it}. The third factor is the autologistic model (4.3). The next factor $p(\boldsymbol{\theta})$ is the prior on the HM-specific parameters $\boldsymbol{\theta}$ in the sampling model. The penultimate factor $p(\boldsymbol{\beta} \mid G)$ is the prior on the non-zero coefficients in the autologistic model. We assume independent normal priors $p(\beta_{ij} \mid G) \propto N(0, 10^6)$. The last factor $p(G)$ is the prior model of Section 3.2.

5 POSTERIOR INFERENCE

We carry out inference for model (4.6) using MCMC posterior simulation. We use $[x \mid y, z]$ to generically indicate a transition probability that modifies x and is indexed by currently imputed values of y and z. The actual transition probability could, for example, be a Gibbs sampling step, replacing the parameter x by a draw from the conditional posterior. MCMC posterior simulation proceeds by iterating over the following transition probabilities:

$$[\boldsymbol{e} \mid G, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{Y}],\ [\boldsymbol{z} \mid \boldsymbol{e}, \boldsymbol{\theta}, \boldsymbol{Y}],\ [\boldsymbol{\theta} \mid \boldsymbol{z}, \boldsymbol{Y}],\ [\boldsymbol{\beta} \mid \boldsymbol{e}, G],\ [G \mid \boldsymbol{\beta}, \boldsymbol{e}]$$

We start by generating e_{it} from its complete conditional posterior. Let $\boldsymbol{e}_{-it} = (e_{ht},\ h \neq i)$ denote the indicators for all HMs other than i at genomic location t and let $i \sim j$ indicate that i and j are neighbors in the graph G. We update e_{it}, $i = 1, \ldots, m$, using

$$p(e_{it} \mid \boldsymbol{e}_{-it}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{Y}) \propto \exp\left\{ \beta_i e_{it} + \sum_{j:\, j \sim i} \beta_{ij} (e_{it} - \nu_i)(e_{jt} - \nu_j) \right\} p(\boldsymbol{Y} \mid \boldsymbol{e}, \boldsymbol{\theta}),$$

and repeat the same loop for each $t = 1, \ldots, n$.

Note that $\boldsymbol{e}_t$, $t = 1, \ldots, n$ are conditionally independent given all other parameters and $\boldsymbol{Y}$. Following the update of $\boldsymbol{e}$ we can then generate new values for $\boldsymbol{z}$ by generating from the complete conditional posterior $p(z_{it} \mid e_{it}, \boldsymbol{\theta}, \boldsymbol{Y})$. If $e_{it} = 0$ the update is deterministic, $z_{it} \equiv -1$. If $e_{it} = 1$ the update requires a Bernoulli draw for $z_{it} = 0$ versus $z_{it} = 1$. The HM-specific parameters $\boldsymbol{\theta}$ are updated by draws from the complete conditional posterior distributions. The details of these transition probabilities are straightforward.

Updating $\boldsymbol{\beta}$ and G is complicated by the fact that the complete conditional posterior distributions for $\boldsymbol{\beta}$ and G require evaluation of the normalization constant that appears in (4.3), given by

$$c(\boldsymbol{\beta}, G) = 1/p(\mathbf{0} \mid \boldsymbol{\beta}, G) = \sum_{\boldsymbol{e}} \exp\left\{ \sum_i \beta_i e_i + \sum_{i<j} \beta_{ij}(e_i - \nu_i)(e_j - \nu_j) \right\}. \qquad (5.1)$$

For later reference let $K(\boldsymbol{e}_t; \boldsymbol{\beta}, G) = \exp\left\{ \sum_i \beta_i e_{it} + \sum_{i<j} \beta_{ij}(e_{it} - \nu_i)(e_{jt} - \nu_j) \right\}$ denote the exponential factor in (4.3), i.e., the un-normalized autologistic conditional probability for e_t. The un-normalized joint probability model for $\boldsymbol{e}$ is thus given as $K(\boldsymbol{e}; \boldsymbol{\beta}, G) = \prod_t K(\boldsymbol{e}_t; \boldsymbol{\beta}, G)$.

The constant $c(\boldsymbol{\beta}, G)$ is computationally intractable for the massively repeated evaluation that is needed in MCMC simulation. The problem of evaluating the normalization constant in the autologistic model is known to be a challenging computational problem. See, for example, Welsh (1990) for a detailed discussion.

We instead implemented an importance sampling estimate to approximate the ratio of the normalizing constants $c(\boldsymbol{\beta}, G)$ required for the evaluation of the acceptance probabilities in the Metropolis-Hastings steps to update β and G. The use of importance sampling estimates to evaluate ratios of normalizing constants is discussed in Chen and Shao (1997), and more recently reviewed in Chen, Shao and Ibrahim (2000, chapter 5). Atchade et al. (2008) discusses the theoretical basis for these algorithms.

We describe the method in detail . Recall that $K(\boldsymbol{e};\ \boldsymbol{\beta}, G)$ is the un-normalized autologistic probability. Let $p_e(\cdot \mid \boldsymbol{\beta}, G)$ denote the autologistic prior model 4.3. We generate a proposal for a new $\tilde{\boldsymbol{\beta}}$ by drawing from $\tilde{\beta}_i \sim q(\tilde{\beta}_i;\ \beta_i) = N(\beta_i, c)$. Next we sample k binary vectors $v_i \sim p_e(v_i;\ \boldsymbol{\beta}, G)$. By the law of large numbers, the sample average $RR = \frac{1}{k}\sum_{i=1}^{k} K(\boldsymbol{v}_i; \tilde{\boldsymbol{\beta}}, G)/K(\boldsymbol{v}_i; \boldsymbol{\beta}, G)$ converges to

$$\int \frac{K(z; \tilde{\boldsymbol{\beta}}, G)}{K(z; \boldsymbol{\beta}, G)} p_e(z; \boldsymbol{\beta}, G) dz = \int \frac{K(z; \tilde{\boldsymbol{\beta}}, G)}{c(\boldsymbol{\beta}; G)} dz = \frac{c(\tilde{\boldsymbol{\beta}}; G)}{c(\boldsymbol{\beta}; G)}.$$

We use RR to estimate the ratio of the normalization constants $c(\tilde{\boldsymbol{\beta}}, G)/c(\boldsymbol{\beta}, G)$ that appears in the Metropolis-Hastings acceptance ratio

$$\frac{p(\tilde{\boldsymbol{\beta}}) K(\boldsymbol{e} \mid \tilde{\boldsymbol{\beta}}) c(\boldsymbol{\beta}, G)}{p(\boldsymbol{\beta}) K(\boldsymbol{e} \mid \boldsymbol{\beta}) c(\tilde{\boldsymbol{\beta}}, G)}$$

for the proposal $\tilde{\boldsymbol{\beta}}$.

In our experience, the importance sampling is fast and sufficiently accurate with an importance sampling size of $k = 5,000$.

We similarly construct another transition probability to update G in a Metropolis-Hastings step. An added complication is that a change in G requires to add or remove coefficients in the autologistic model (4.3), thus changing its dimension. Consequently, a transition probability for G requires trans-dimensional MCMC. We use a reversible jump (RJ)-MCMC (Green, 1995). We construct a candidate $\widetilde{G}$ by adding or deleting an edge from G.

6 SIMULATION STUDY

6.1 Uniform Prior on a Small Graph

We evaluate the proposed model and inference approach with 10 simulated data sets of hypothetical observations of 10 HMs at $n = 8,000$ genomic locations. For each of the 10 simulations, we first generated a simulation truth G_1 by setting up vertices for $m = 10$ variables. For each pair of vertices $\{i, j\}$ we included an edge between them with probability 0.25. In other words, the simulation truth is generated from a uniform prior for G, uniform over all possible G with 10 vertices for a given graph size $|E|$, and a binomial prior on $|E|$ determined by the probability 0.25 of including each possible edge. For each imputed edge $\{i, j\}$ we generated values of β_{ij} in (4.3) using a uniform prior over three possible values, $\beta_{ij} \sim$ Unif($\{\log(2), \log(4), -\log(2)\}$. These values were chosen arbitrarily. Autologistic intercepts β_i in (4.3) were generated as $\beta_i \sim N(0, .3)$. We then generated $\boldsymbol{e}$ from the autologistic prior model (4.3) using a Gibbs sampler simulation with $8,000$ iterations. We kept the last draw as the simulation truth for $\boldsymbol{e}$. Conditional on $\boldsymbol{e}$ we then imputed hypothetical HM counts $\boldsymbol{Y}$ from (4.4), using the following HM-specific parameters $\boldsymbol{\theta}$. We fixed $\sigma_{1i} = \sigma_{2i} = .2$ and generated $\mu_{1i} \sim N(log(4), .25)$ and $\mu_{2i} \sim N(log(5), .25)$. The thresholds c_i were fixed at 3. The Poisson rates were generated as $\lambda_i \sim N(1, .1)$.

For each simulated data set we implemented the described posterior MCMC simulation to compute posterior summaries. The posterior estimates are computed using MCMC posterior simulation with an initial burn-in of 7,000 iterations and a total of 10,000 iterations for all 10 simulations. We compared inference for G under the proposed model versus the simulation truth for each of the 10 simulated datasets. We took the $k = 3,000$ post burn-in MCMC samples and computed the posterior inclusion probability $\widehat{P}_{ij}$ for each possible edge $\{i, j\}$ in

the graph, which is defined as

$$\widehat{P}_{ij} = \frac{1}{k} \sum I(\{i, j\} \in E),$$

substituting the edge set E of the imputed graph for each iteration of the MCMC. We then estimate the graph G by including all edges with $\widehat{P}_{ij} > 0.5$. This estimate is also known as the median inclusion probability model (Barbieri and Berger, 2004). For 5 of the simulations the posterior estimated graph exactly matched the simulation truth, and for the remaining 5 simulations they differed by up to two edges.

Next, we report parameter estimates. Let $\boldsymbol{\beta}^o$ denote the simulation truth, including $\beta_{ij}^o = 0$ for edges that are not included in the simulation truth for G. Let $\overline{\boldsymbol{\beta}}^{(s)} = E(\boldsymbol{\beta} \mid \boldsymbol{Y}^{(s)})$ denote the posterior mean for the autologistic coefficients conditional on the s-th simulated data set $\boldsymbol{Y}^{(s)}$. Here $\boldsymbol{Y}^{(s)}$ are the simulated HM counts, for $s = 1, \ldots, 10$. In the posterior mean we evaluate β_{ij} as $\beta_{ij} = 0$ when the corresponding edge is missing in the imputed graph G. We calculated the mean squared errors (MSE) as $\text{MSE}_{ij} = (1/10) \sum_{s=1}^{10} \left(\overline{\beta}_{ij}^{(s)} - \beta_{ij}^o \right)^2$. Similarly we compute MSE for the posterior means $\bar{\boldsymbol{\theta}}^{(s)}$ of the parameters that index the sampling model (4.4). Table 1a lists some MSE values. Recall that the simulation truth for all parameters was around 1.0 in absolute values, leaving the reported MSE values quite acceptable. The numerical errors of the MSE values are reported in parentheses besides the reported MSE estimates in Table 1a.

Results in Table 1a are based on a uniform prior over G. We explored sensitivity with respect to the prior choice for $p(G)$ by repeating the same simulation with an alternative informative prior. We used (3.1) with G_0 specified as a slight variation of the simulation truth which we achieved by continuing MCMC iteration for 500 iterations. To compare the simulation truth and the estimated graph we compute the proportion of mismatched edges in the posterior estimate. We evaluated average error rates over 10 simulations each for

a variety of combinations of sample size n and prior hyper-parameter ρ. For sample sizes $n > 800$ the value of ρ does not seem to affect the accuracy rates. For $n = 850$ we find error rates between 0.044 and 0.015 for ρ between 0.50 through 0.95. Only as the sample size falls below $n = 450$ we find that we need values of $\rho > 0.95$ to achieve reasonable accuracy rates. In the simulation we found average error rates between 0.122 and 0.066 for ρ between 0.85 and 0.99.

6.2 Informative Prior for a Large Graph

We tested the performance of the uniform prior for larger graphs with $m = 39$ nodes. With the same simulation setup as in Section 6.1, we find large error rates of 20% and more, leading us to advise against the use of uniform priors for larger networks. This observation is in agreement with similar advise in the recent literature (Jones et al., 2004; Ellis and Wong, 2008). Instead, we investigate the performance of the informative prior (3.1) for a graph with all 39 HMs. We keep all the simulation settings unchanged except that we decrease the probability of generating an edge between any pair of vertices, $\{i, j\}$, from 0.25 to 0.05, lest estimated graphs become impossible to display and interpret. We also increased the number of hypothetical loci to 45,000.

As prior $p(G)$ we use the informative prior (3.1). We center the prior at a network G_0 that is constructed to be a distorted version of the simulation truth. We generate G_0 by continuing prior simulation that was used to generate the simulation truth G_1 for 500 more MCMC iterations. The construction is meant to mimic the nature of an informative prior based on expert opinion or related data.

As in the first simulation, we randomly generate 10 datasets and estimate the graphical models to each of the 10 datasets. The posterior estimates are computed using MCMC posterior simulation with a total of 5,500 iterations and an initial burn-in of 3,500 iterations.

We found that the number of burn-ins are adequate by inspection of the trace plots. We verified practical convergence using Geweke's convergence diagnostic (Geweke, 1991) and found no indication of lack of convergence.

We take the 2,000 post burn-in MCMC samples and compute the posterior inclusion probability $\widehat{P}_{ij}$ for each possible edge $\{i, j\}$ in the graph. We then estimate the graph G by including all edges with $\widehat{P}_{ij} > c$ where the threshold c is chosen to achieve a posterior expected false discovery rate (FDR) close to .01. The posterior expected FDR for any given threshold c is calculated by

$$FDR_c = \frac{\sum_{i\,j} [(1 - \widehat{P}_{ij}) I(\widehat{P}_{ij} > c)]}{\sum_{i,j} I(\widehat{P}_{ij} > c)}$$

To compare the simulation truth and the estimated graph, we compute for each simulation an error rate

$$ER = 1 - (|(E_1 \cap \widehat{E})/|E_1|) \tag{6.1}$$

where $|\cdot|$ is the cardinality of a set, E_1 is the edge set of the simulation truth, and $\widehat{E}$ is the edge set of the estimated posterior graph. We find an average ER of 0.09 over the 10 simulations. Figure 2 compares the estimated graphs to the true graph for one simulated dataset. Table 1b lists the MSE values for selected $\boldsymbol{\beta}$ parameters.

6.3 Comparisons with Gaussian Graphical Models

We compared inference under the proposed graphical model with inference under Gaussian graphical models (GGMs), using the graphical lasso method of Friedman et al. (2008) to implement inference under the GGM. We used the R package *g-lasso*. HM counts for 10 hypothetical datasets were generated from the autologistic model described in Section 6.1. Following the simulation setup in Friedman et al. (2008), we set the penalization parameter λ for each dataset such that the estimated number of non-zero entries is closest to the number

of non-zero entries in the simulation truth, noting that this choice favors the g-lasso. The resulting λ values for the 10 imputed data sets were between 0.022 and 0.035. For each data set we computed realized error rates ER as defined in (6.1). Averages across the 10 simulations are summarized in Table 2. When the true model is autologistic, the error rate for the *g-lasso* estimate is substantially higher than under the proposed model. For the HM data we believe that the autologistic model better reflects the underlying biology since biologically, events such as modifications of histones are truly binary, which are modeled as our latent binary indicators. The results of Table 2 would thus favor inference with the proposed model for HMs. In general, the proposed model is applicable for the analysis of data having latent binary structures.

7 ChIP-SEQ DATA ANALYSIS

7.1 Inference for 17 HMs

Wang et al (2008) suggest that multiple HMs can influence critical regulatory elements of transcription in a combinatorial fashion, To identify HMs that may function together to modify chromatin, they searched for robust modification features at promoter regions. This analysis revealed an HM "backbone" consisting of 17 HMs. We implemented inference in the proposed model for the ChIP-Seq data for this selected subset of $m_1 = 17$ HMs, using the uniform graphical prior $p(G)$. The names of the 17 HMs and their ids are provided in the Appendix.

We considered inference for global dependence structure, i.e., inference for one graphical model G that represents conditional independence across all loci. Note that one common dependence structure across all loci does not exclude varying abundance of HM for different types of genomic regions. Later, in the next section, we will alternatively also consider

models for varying dependence structure.

We carried out inference using the described MCMC posterior simulation. For purposes of this analysis, we randomly sampled 50,000 genomic locations from the pre-processed dataset. For each edge $\{v_i, v_j\}$ we obtained the posterior inclusion probability $\widehat{P}_{ij}$, defined as before. Similarly, posterior estimates of the parameters $\boldsymbol{\theta}$ that index the sampling model were obtained by averaging over the 3,000 posterior samples, after discarding the initial $7,000$ iterations.

We obtained a posterior estimate of the unknown graph G for the selected 17 HMs by connecting any pair of vertices with posterior inclusion probability $\widehat{P}_{ij} > .95$. The threshold was chosen to achieve a posterior expected false discovery rate (FDR) close to .01.

The estimated graph is shown in Figure 3. The colors (and line types) of the edges were determined by $p(\beta_{ij} > 0 \mid \beta_{ij} \neq 0, \boldsymbol{Y})$. If the reported probability is greater than $1/2$, then we say that the variables connected by the edge have an enhancing, positive relationship. Otherwise we say that they have an inhibiting, negative relationship. We denote the positive and negative edges by solid (blue) lines and dashed (red) lines, respectively.

We observe that the posterior graph in Figure 3 is a highly connected and dense graph. In terms of the properties of the joint distribution, it implies that modification types are marginally highly correlated.

In Figure 4, we show the four most frequent configurations of a set of 4 edges, based on posterior inclusion probabilities. This set comprises of the edges shown in Figure 4(d). As expected, the configuration of the 4 edges that appears in Figure 3 is the most frequent one and is shown in Figure 4(a).

Finally, as a benefit of full posterior inference, we obtain posterior probabilities of the presence or absence of HMs. These are biologically meaningful parameters. The Bayesian network models the HMs as continuous variables and fails to provide any inference on the latent biological signal.

7.2 Biological Annotations for the 17 HM Backbone Network

In the analysis of the 17 backbone HMs, strong positive edges are observed between the following pairs in Figure 3

H2BK5ac, H3K27ac	H3K4ac, H3K91ac	H2Az, H3K4ME3
H4K91ac, H2BK120ac	H4K91ac, H2BK20ac	H4K8, H4K5ac

These are the set of edges with values of $\hat{\beta}_{ij} > .8$, where $\hat{\beta}_{ij}$ is the posterior mean of β_{ij}

The same pairs have high pairwise correlation in the heat-map analysis reported by Wang et al. (2008, Figure 4 therein). Some of these results validate the related findings that are reported in the literature on HMs. For example, it is well known that H3K4me3 modification is strongly correlated with active transcription and often co-localized with H2A.Z enrichment (Barski et al. 2007). In Bártfai et al. (2010) they find evidence of an almost perfect co-localization of H2A.Z with H3K9ac and H3K4me3 in the plasmodium genome. The authors suggest that these marks are preferentially deposited on H2A.Z-containing nucleosomes. Karlić et al. (2010) mention H2BK5ac and H3K27ac in a list of 4 HMs that appear to be the most important modifications associated with gene expression levels. Of these, H2BK5ac and H3K27ac had the highest individual information content and their levels were highly correlated. They went on to suggest that these two HMs are the most important ones for gene expression. Interestingly, positive associations are also obtained between H3K4 and H3K9 methylation groups. We shall refer to this again while discussing the results on 39 HMs.

7.3 Full Data of 39 HMs

The previous analysis assumed one common dependence structure across the entire genome. We now relax this assumption and focus on a subset of the ChIP-seq data covering the 33,681 loci in promoter regions. Among these loci, 3,895 are close to non-coding RNA and 29,786

are close to protein coding genes. We further focus on the regions close to coding regions and classify them as high versus low expression. This data allows us to study differences in dependence structure between regions that correspond to high and low gene expressions.

We define high and low expression by considering data (Su et al., 2004) on the expression levels of the corresponding genes. Using the 25th and 75th percentiles as the threshold, we define the subsets of the promoter regions corresponding to low expression ($<$ 25th percentile) and high expression ($>$ 75th percentile).

We use HM counts from the excluded non-coding regions to construct an informative prior. Specifically, we use the prior model (3.1) with the centering graph G_0 defined by the estimation of a Bayesian network for the non-coding regions. Figure 5 presents the prior centering graph G_0, and the posterior estimated graphs correspond to the regions with high and low expression, respectively. For ease of display, we used a numeric id for each HM (instead of its full name) in plotting the graph. Their names are given in Appendix. Table 3 gives a measure of the similarity between the HM networks in the regions of high and low gene expression. For each type of network E_1, a non-matching rate (NMR) relative to the respective other network E_2 is calculated as the proportion of edges in E_1 that are not present in E_2. The proportion is relative to the size of E_1. That is, NMR is $1-|E_1 \cap E_2|/|E_1|$. Here E_1 and E_2 refer to the two networks for coding regions corresponding to low and highly expressed genes, respectively. Matching rate (MR) is simply 1-NMR. The NMR values that are reported in Table 3 for the high and low expression datasets indicate that the networks in these two regions are highly dissimilar.

The network for the high expression regions is denser than that in the low expression regions, which is denser than the prior network G_0 for the non-coding region. This finding agrees with what is discovered in Rosenfeld et al. (2009). However, there are some edges that occur with high posterior probability in both these networks; these common edges include (H3K14ac, H4K20me1), (H3K27me1,H2BK20ac), (H3K79me3, H3K9me3), as well as H3K27

and H3K9 methylation groups.

In the high expression network in Figure 5(b), strong positive edges are observed between pairs connecting H3K9ac and H3K27me3, H3K4me3 and H3K20me3; and some more pairs.

In the low expression network in Figure 5(c), we observe strong positive associations between pairs such as H3K4me1 and H3K9me1.

Lastly we list the top HMs that have high posterior probability for high connectivity, defined as having five or more edges. In the high expression network, the top HMs include H3K36me3, H4K20me3, H3K27ac, and H3K9ac. In the low expression network, the prominent HMs are H3K9me1 and H3K27AC. Barski et al. (2007) find that the H3K9 methylations play a significant role in gene repression. So we expect it to play a prominent role in the network for low gene expressions. We indeed find that it is the most connected node in the low expression network.

Some of these findings require further studies and validations. An interesting example is the positive connection between H3K4 and H3K9 methylations.

This may seem at first sight paradoxical because H3K4ME3 is known to be associated with gene activation whereas H3K9ME3 is associated with gene repression. However, activating H3K4ME3 and repressive H3K27ME3 are known to be over represented in the promoters of embryonic stem cells. These bivalent genes are either to be activated or silenced upon differentiation. Therefore, activating and repressive HM marks occupying the same regions are potentially interesting biologically. In fact, consistent with our finding, it has been suggested that H3K4ME3 and H3K9ME3 co-marked open reading frames are engaged in dynamic transcriptional activity (Berger 2007). H3K4ME3 and H3K9ME3 are found to be associated with different HMs in low and high expression sets (Figure 5). The activating H3K4ME3 is associated with four acetylation marks as well as H3K36ME1 in high expression set. All six HMs are activating. In low expression set, H3K4me3 is associated with H3K4me1 and H3R2me1 which are less activating. H3K9me3 in high expression set

is associated with activating H2AK9ac and H3K36me1, whereas in the low expression set H3K9me3 is associated with repressive H3K27me3. This example indicates that the findings of our model are biologically reasonable.

8 DISCUSSION

The proposed Bayesian graphical model implements an approach to decode the dependence structure of HMs. Graphical models are not new and have been previously applied to biological networks. We build on previous approaches and generalize them in two important directions to adapt them for the HM problem. We employed Markov random fields to specify the relationships among binary indicators for HMs. By this, we are able to build a model that includes positive prior probability for all possible forms of conditional independence structures among a selected set of HMs. The integration of the autologistic model and latent variable modeling in a Bayesian framework is novel. Presence and absence of HMs are modeled through hidden binary variables while the autologistic framework modeled the relationships between them. Though introduced decades earlier in Besag (1974), the scope of application of autologistic models has been limited due to computationally intractable normalization constants in the model. We get around this computational challenge by using an importance sampling approximation. In the application to HMs, we employ a centered version of the model that greatly improves mixing of posterior MCMC simulations. The combined use of these model choices and techniques enables us to report joint inference about the latent signals of HMs and their underlying dependence.

Our results are suggestive of a crosstalk mechanism between the HMs. The posterior graph is shown to be highly connected. Most inference confirmed and quantitatively evaluated known relationships. In a few cases, however, the strength of the association do not match the hypothesized relationships. This is due to a number of factors. One potential

limitation is that we are modeling the HM dependence across all types of genomic locations. In the case of differential histone patterns, the model would report only associations that are strong and universally present across all the locations. In future research we will investigate formal modeling of changing dependence patterns. A related natural next step is inference for differential histone patterns, comparing dependence structure across biological conditions, using one encompassing model that includes biologic condition as a covariate.

References

Andersson, R., Enroth, S., Rada-Iglesias, A., Wadelius, C., and Komorowski, J. (2009). Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Res.*, 19:1732–1741.

Atchade, Y., Lartillot, N., and Robert, C. (2008). Bayesian computation for statistical models with intractable normalizing constants. *Technical report, University of Michigan, Department of Statistics*.

Barbieri, M. and Berger, J. (2004). Optimal predictive model selection. *The Annals of Statistics*, 32(3):870–897.

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, 129:823–837.

Bártfai, R., Hoeijmakers, W. A. M., Salcedo-Amaya, A. M., Smits, A. H., Janssen-Megens, E., Kaan, A., Treeck, M., Gilberger, T.-W., Françoijs, K.-J., and Stunnenberg, H. G. (2010). H2a.z demarcates intergenic regions of the plasmodium falciparum epigenome that are dynamically marked by h3k9ac and h3k4me3. *PLoS Pathog*, 6(12):e1001223.

Bergink, S., Salomons, F. A., Hoogstraten, D., Groothuis, T. A., de Waard, H., Wu, J., Yuan, L., Citterio, E., Houtsmuller, A. B., Neefjes, J., Hoeijmakers, J. H., Vermeulen, W., and Dantuma, N. P. (2006). DNA damage triggers nucleotide excision repair-dependent monoubiquitylation of histone H2A. *Genes Dev.*, 20:1343–1352.

Bernstein, B. E., Humphrey, E. L., Erlich, R. L., Schneider, R., Bouman, P., Liu, J. S., Kouzarides, T., and Schreiber, S. L. (2002). Methylation of histone H3 Lys 4 in coding regions of active genes. *Proc. Natl. Acad. Sci. U.S.A.*, 99:8695–8700.

Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems . *Journal of Royal Statistical Society Series B*, 135:192–236.

Caragea, C. and Kaiser, S. (2009). Autologistic models with interpretable parameters. *Journal of Agricultural, Biological, and Environmental statistics*, 14:281–300.

Chen, M.-H. and Shao, Q.-M. (1997). On Monte Carlo methods for estimating ratios of normalizing constants. *The Annals of Statistics*, 25:1563–1594.

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer, Verlag New York.

Ellis, B. and Wong, W. H. (2008). Learning Causal Bayesian Network Structures From Experimental Data. *Journal of the American Statistical Association*, 103:778–789.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Geweke, J. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*. Federal Reserve Bank of Minneapolis, Research Dept.

Grant, P. A. and Berger, S. L. (1999). Histone acetyltransferase complexes. *Semin. Cell Dev. Biol.*, 10:169–177.

Green, R. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82.

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., Ye, Z., Lee, L. K., Stuart, R. K., Ching, C. W., Ching, K. A., Antosiewicz-Bourget, J. E., Liu, H., Zhang, X., Green, R. D., Lobanenkov, V. V., Stewart, R., Thomson, J. A., Crawford, G. E., Kellis, M., and Ren, B. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459:108–112.

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2004). Experiments in Stochastic Computation for High-Dimensional Graphical Models. *Statistical Science*, 20:388–400.

Kalisch, M. and Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *The Journal of Machine Learning Research*, 8:613–636.

Karlić, R., Chung, H.-R., Lasserre, J., Vlahovicek, K., and Vingron, M. (2010). Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci U S A*, 107(7):2926–31.

Kim, T. H., Barrera, L. O., Zheng, M., Qu, C., Singer, M. A., Richmond, T. A., Wu, Y., Green, R. D., and Ren, B. (2005). A high-resolution map of active promoters in the human genome. *Nature*, 436:876–880.

Kuan, P. F., Chung, D., Pan, G., Thomson, J. A., Stewart, R., and Keles, S. (2011). A statistical framework for the analysis of chip-seq data. *Journal of the American Statistical Association*, 106(495):891–903.

Liu, C. L., Kaplan, T., Kim, M., Buratowski, S., Schreiber, S. L., Friedman, N., and Rando,

O. J. (2005). Single-nucleosome mapping of histone modifications in S. cerevisiae. *PLoS Biol.*, 3:e328.

Marks, P., Rifkind, R. A., Richon, V. M., Breslow, R., Miller, T., and Kelly, W. K. (2001). Histone deacetylases and cancer: causes and therapies. *Nat Rev Cancer*, 1(3):194–202.

Parmigiani, G., Garrett, S., Anbazhagan, R., and Gabrielson, E. (2002). A Statistical Framework for Expression-Based Molecular Classification in Cancer. *Journal of Royal Statistical Society Series B*, 64:717–736.

Pokholok, D. K., Harbison, C. T., Levine, S., Cole, M., Hannett, N. M., Lee, T. I., Bell, G. W., Walker, K., Rolfe, P. A., Herbolsheimer, E., Zeitlinger, J., Lewitter, F., Gifford, D. K., and Young, R. A. (2005). Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell*, 122:517–527.

Roh, T. Y., Cuddapah, S., and Zhao, K. (2005). Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev.*, 19:542–552.

Rosenfeld, J., Wang, Z., Schones, D., Zhao, K., DeSalle, R., and Zhang, M. (2009). Determination of enriched histone modifications in non-genic portions of the human genome. *BMC genomics*, 10(1):143.

Rozowsky, J., Euskirchen, G., Auerbach, R., Zhang, D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M., and Gerstein, M. (2009). PeakSeq enables systematic scoring of ChIP-Seq experiments relative to controls. *Nature Biotechnology*, 27:66–75.

Strahl, B. D. and Allis, C. D. (2000). The language of covalent histone modifications. *Nature*, 403:41–45.

Su, R.-C., Brown, K. E., Saaber, S., Fisher, A. G., Merkenschlager, M., and Smale, S. T.

(2004). Dynamic assembly of silent chromatin during thymocyte maturation. *Nat Genet*, 36(5):502–6.

Telesca, D., Muller, P., Kornblau, S., and Ji, Y. (2010). Modeling protein expression and dependent pathways. *Technical Report, UT MD Anderson Cancer Centre*.

Wang, Z., Zang, C., Rosenfeld, J. A., Schones, D. E., Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Peng, W., Zhang, M. Q., and Zhao, K. (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.*, 40:897–903.

Welsh, D. J. A. (1990). The computational complexity of some classical problems from statistical physics. In *In Disorder in Physical Systems*, pages 307–321. Clarendon Press.

Zang, C., Schones, D. E., Zeng, C., Cui, K., Zhao, K., and Peng, W. (2009). A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics*, 25:1952–1958.

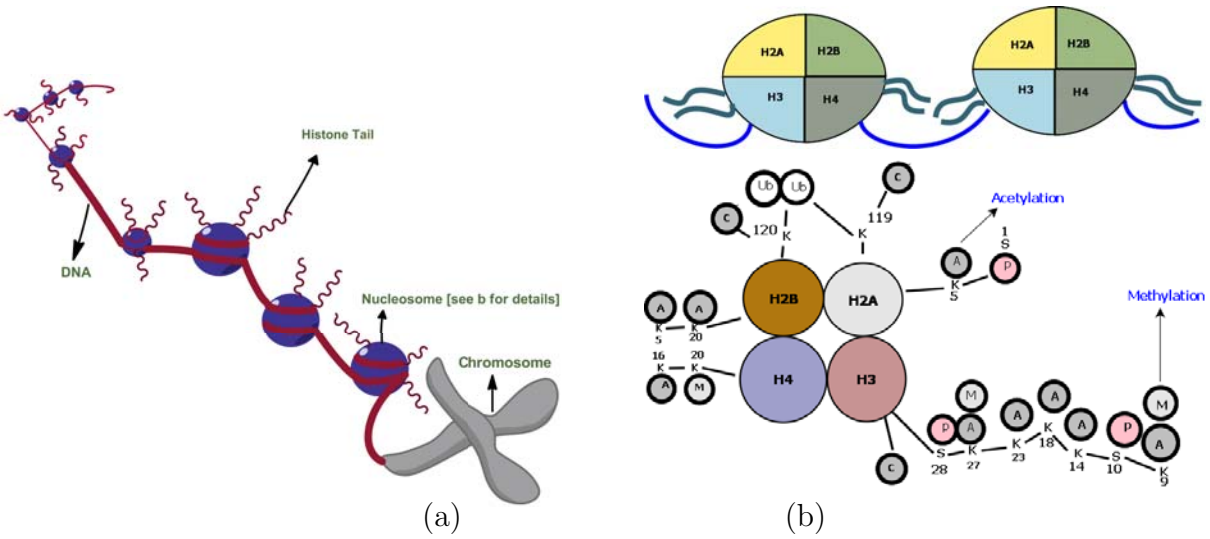
Zhang, Y., Lv, J., Liu, H., Zhu, J., Su, J., Wu, Q., Qi, Y., Wang, F., and Li, X. (2010). HHMD: the human histone modification database. *Nucleic Acids Res.*, 38:D149–154.

APPENDIX: LIST of the HMs

Below are two sets of index of HMs selected from Wang et al. (2008). The 17-HM index is used in the Figures 3 and 4, and the 39-HMs index is used in Figures 2 and 5.

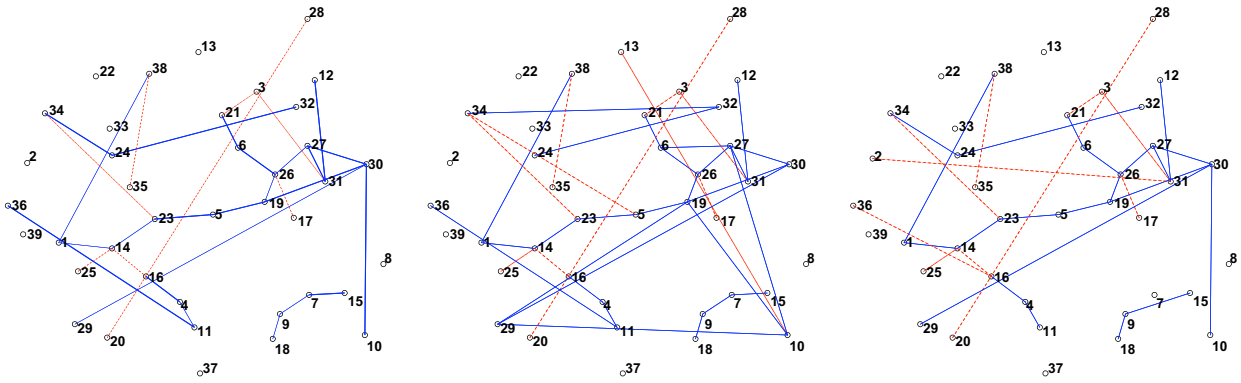
Index	HM
1	H2BK120ac
2	H2BK12ac
3	H2BK20ac
4	H2BK5ac
5	H3K4AC
6	H3K4ME1
7	H3K4ME2
8	H3K4ME3
9	H3K9AC
10	H3K9ME1
11	H3K27AC
12	H3K36AC
13	H3K18AC
14	H4K91AC
15	H2AZ
16	H4K5AC
17	H4K8AC

Index	HM
1	H2AK5ac
2	H2AK9ac
3	H2BK120ac
4	H2BK12ac
5	H2BK20ac
6	H2BK5ac
7	H3K14AC
8	H3K18AC
9	K23AC
10	H3K27AC
11	H3K36AC
12	H3K4AC
13	H3K9AC
14	H4K12AC
15	H4K16AC
16	H4K5AC
17	H4K8AC
18	H4K91AC
19	H2AZ
20	H3BK5ME1
21	H3K27ME1
22	H3K27ME2
23	H3K27ME3
24	H3K36ME1
25	H3K36ME3
26	H3K4ME1
27	H3K4ME2
28	H3K4ME3
29	H3K79ME1
30	H3K79ME2
31	H3K79ME3
32	H3K9ME1
33	H3K9ME2
34	H3K9ME3
35	H4R2ME1
36	H4R2ME2
37	H4K20ME1
38	H4K20ME3
39	H4R3ME2



(a) (b)

Figure 1: Panel (a) illustrates how DNA is wrapped around nucleosomes to form a picture like beads on a string. Panel (b) focuses on one of the nucleosomes. Nucleosomes consist of four core histone (H) proteins, H2A, H2B, H3, and H4. Each histone has two copies per nucleosome, thus forming an octamer of histones. DNA are densely wrapped around these octamers forming *chromatins*. Lysines (K) in the amino-terminal tails of histones H2A, H2B, H3 and H4 are potential sites for modifications, including acetylation (A), phosphorylation (P), methylation (M), ubiquitination (Ub), and sumoylation (Su). For example, H3K9ac stands for an acetylation at the residue 9 (which is a lysine) of histone H3. The figure is simplified from Marks et al. (2001).



(a) Simulation truth (b) Posterior estimate (uniform) (c) Posterior estimate (informative)

Figure 2: The simulated truth versus the estimated graph of one simulated dataset based on 39 vertices and the informative prior for the random graph. The concentration parameter ρ is set to 0.93. In (a) we present the simulation truth. In (b) we present the posterior estimated graph under a uniform prior. In (c) we present the posterior estimated graph when the informative prior is centered at a graph that is slightly and randomly perturbed from the simulation truth. The dotted (red) lines represent the HM pairs with negative relationships, while the solid (blue) lines depict the positive ones.

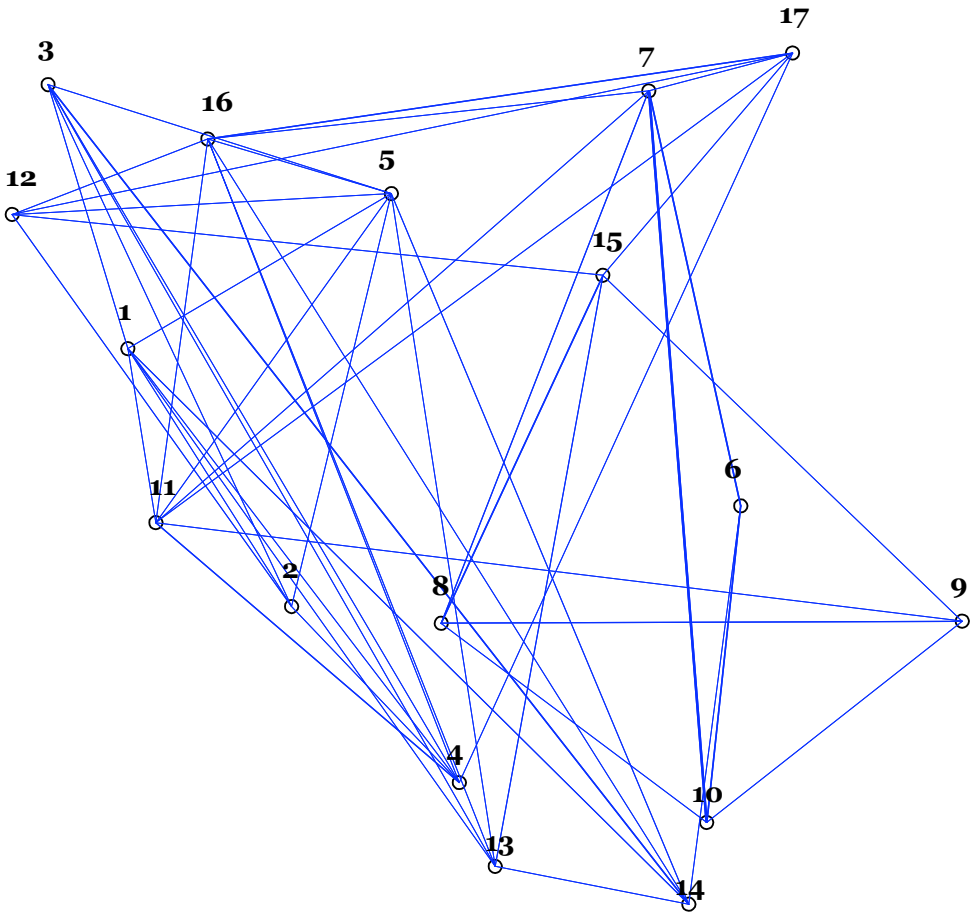
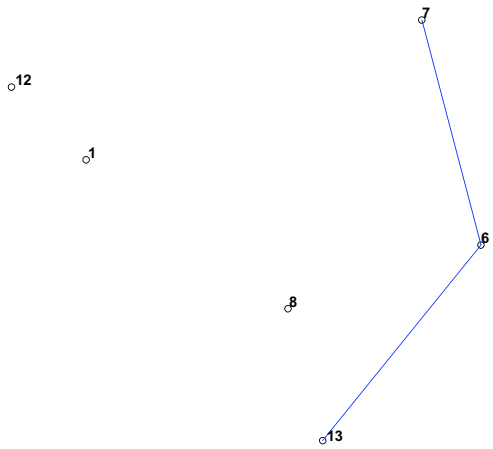
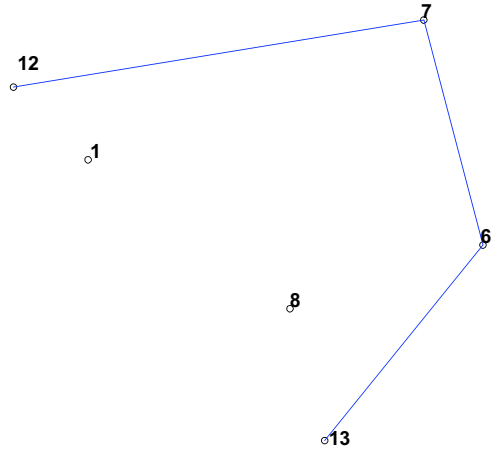


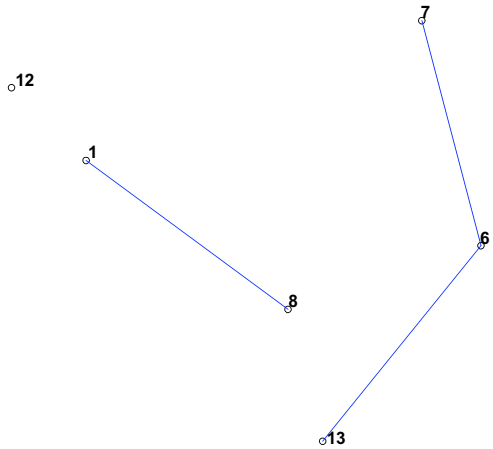
Figure 3: Posterior inference of the ChIP-Seq data based on 17 HMs (see Appendix) and a uniform prior.



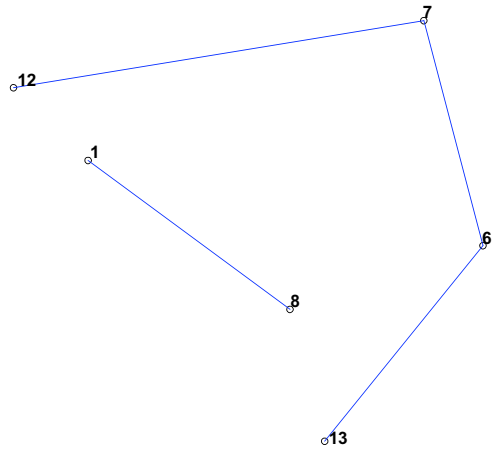
(a) 55



(b) 15

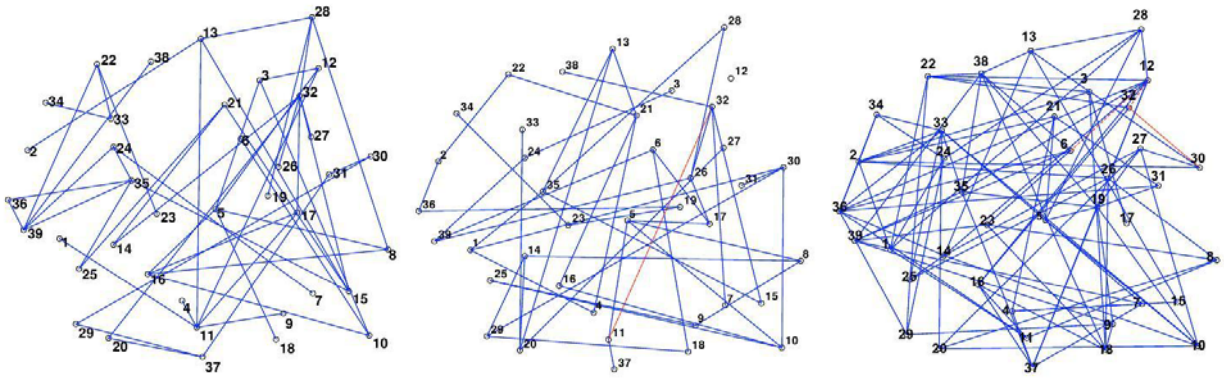


(c) 13



(d) 12

Figure 4: The four most frequent configurations (a through d) of a subgraph consisting of 4 edges. The posterior probabilities (in percent) are given below each subgraph.



(a) Prior (b) Low expression (c) High expression

Figure 5: Posterior inference of the ChIP-Seq data based on 39 HMs and informative prior. In (a), we present the estimated HM network for non coding regions, which used as a prior centering graph for inference in regions of high and low gene expression. In (b) and (c), we present posterior estimated graphs for regions of low and high expression, respectively.

Table 1: Simulation study: MSE values of the parameters and their standard errors (SE). We arbitrarily selected four $\boldsymbol{\beta}_{ij}$ parameters that had non-zero values in the simulation truth.

(a) small graph

Parameter	MSE
λ_1	.03(.008)
σ_{17}	.04(.009)
σ_{21}	.05(.010)
μ_{17}	.02(.004)
μ_{21}	.02(.005)
$\beta_{2,10}$	.04(.021)
$\beta_{8,5}$	.08(.037)
$\beta_{4,9}$	.06(.028)
$\beta_{6,1}$	.07(.033)

(b) large graph

Parameter	MSE
$\beta_{15,4}$	.05(.005)
$\beta_{22,17}$	.03(.006)
$\beta_{1,33}$	.06(.008)
$\beta_{16,7}$	.03(.005)

Table 2: Average error rates under the GGM and the proposed autologistic model.

Inference Model	Simulation Model	
	GGM	Autologistic
GGM	.05	.14
Autologistic	.12	.03

Table 3: Summary statistics for the three networks. Here NMR is the non-matching rate of one network (E_1) relative to another network (E_2), calculated as the proportion of edges in E_1 that are not present in E_2. In the table below, E_1 and E_2 refer to the two networks for coding regions corresponding to highly and lowly expressed genes, respectively. Matching rate (MR) is simply 1-NMR.

	NMR	MR
High expression	.88	.12
Low expression	.92	.08