The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1749–1772
https://doi.org/10.1214/17-AOAS1129

NONPARAMETRIC BAYESIAN LEARNING OF HETEROGENEOUS DYNAMIC TRANSCRIPTION FACTOR NETWORKS

BY XIANGYU LUO[1] AND YINGYING WEI[2]

The Chinese University of Hong Kong

Gene expression is largely controlled by transcription factors (TFs) in a collaborative manner. Therefore, an understanding of TF collaboration is crucial for the elucidation of gene regulation. The co-activation of TFs can be represented by networks. These networks are dynamic in diverse biological conditions and heterogeneous across the genome within each biological condition. Existing methods for construction of TF networks lack solid statistical models, analyze each biological condition separately, and enforce a single network for all genomic locations within one biological condition, resulting in low statistical power and misleading spurious associations. In this paper, we present a novel Bayesian nonparametric dynamic Poisson graphical model for inference on TF networks. Our approach automatically teases out genome heterogeneity and borrows information across conditions to improve signal detection from very few replicates, thus offering a valid and efficient measure of TF co-activations. We develop an efficient parallel Markov chain Monte Carlo algorithm for posterior computation. The proposed approach is applied to study TF associations in ENCODE cell lines and provides novel findings.

1. Introduction. For a single person, the same copy of a genome can give rise to hundreds of distinct cell types [Lan et al. (1997), Lara-Marquez et al. (2001)]. To understand this phenomenon, functional genomics aims to reveal gene regulation mechanisms. Gene regulation is largely controlled by a family of proteins called transcription factors (TFs), which bind to specific DNA sequences to either activate or repress the expression levels of nearby genes [Mitchell and Tjian (1989)]. Accurate gene expression regulation requires extensive collaborations among TFs [Chen et al. (2012), Hobert (2008)]. As a result, TF cooperation plays a critical role in diverse human diseases, such as renal disease [Zhou et al. (2008)], Parkinson disease [Scherzer et al. (2008)], Alzheimer disease [Kitamura et al. (1997)], and pancreatic cancer [Shi et al. (1999)]. For example, in pancreatic cancer, the collaboration between two TFs—AP-1 and NFκB—regulates the expression of

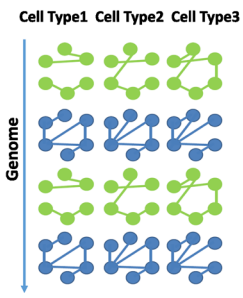
Received March 2017; revised November 2017.

[1]Supported by Hong Kong Ph.D. Fellowship (PF13-11656).

[2]Supported in part by Early Career Scheme 24301416 from the Research Grants Council of the Hong Kong Special Administrative Region and Direct Grants from the Research Committee of the Chinese University of Hong Kong.

Key words and phrases. Poisson graphical model, nonparametric Bayes, parallel Markov chain Monte Carlo, next generation sequencing.

(a) (b)



Cell Type 1						Cell Type 2						Cell Type 3					
TF1		TF2		...		TF1		TF2		...		TF1		TF2		...	
Rep 1	Rep 2	Rep 1	Rep 2	Rep 1	Rep 2	Rep 1	Rep 2	Rep 1	Rep 2	Rep 1	Rep 2	Rep 1	Rep 2	Rep 1	Rep 2	Rep 1	Rep 2
1	4	2	1	...	...	2	5	3	4	...	...	2	6	2	2	...	...
2	6	2	2	...	...	2	2	3	4	...	...	4	4	6	5	...	...
3	3	1	4	...	...	3	2	2	4	...	...	4	3	5	2	...	...
3	3	2	2	...	...	3	2	2	2	...	...	4	1	5	2	...	...

FIG. 1. (a) *Cartoon illustration of TF networks. Each node represents one TF, and each edge indicates the co-activation of a pair of TFs. Each row corresponds to one genomic location. On the one hand, the TF networks vary across cell types. On the other hand, genomic locations with the same color share the same dynamic network. Thus, the TF networks are heterogeneous across the genome.* (b) *Data structure of the observed ChIP-seq counts. Each row corresponds to one genomic location in Figure* 1(a). *Two replicates are measured for each TF under each cell type. The goal is to infer the underlying heterogeneous dynamic TF networks in Figure* 1(a) *from the observed data in Figure* 1(b).

interleukin-8 (IL-8), whose secretion is directly associated with the cancer progression [Shi et al. (1999)]. Therefore, an understanding of TF cooperation is crucial for the elucidation of gene regulation and ultimately disease mechanisms. The interactions of TFs at each genomic location can be represented by a network. Each node of the network corresponds to one TF, and each edge characterizes the dependence of the binding intensities for a pair of TFs. Because TF collaborations differ among biological conditions (or cell types), the network at each genomic location is dynamic. Moreover, as TF interactions change across the genome, all dynamic networks are heterogeneous across the genome rather than sharing the same structure. Consequently, studying TF associations translates into learning genome-wide dynamic TF networks [see Figure 1(a)].

Count data for inferring TF networks can be obtained from ChIP-seq experiments, a technology that couples chromatin immunoprecipitation with high-throughput sequencing [Johnson et al. (2007)]. Although each ChIP-seq experiment can measure the genome-wide binding intensities only for a given TF under a single condition, ChIP-seq data accumulate rapidly in public data repositories. The Encyclopedia of DNA Elements (ENCODE) project [ENCODE Project Consortium (2012)] has collected more than 1200 TF ChIP-seq datasets, which provides an unprecedented opportunity for systematic investigation of TF interactions. However, from the ENCODE project, a typical ChIP-seq experiment has only two to four replicates [see Figure 1(b)]. Consequently, the analysis of TF interactions

separately at each genomic location under each biological condition suffers from very low statistical power.

To improve graph inference, a large body of research has recently been devoted to joint estimation of multiple graphical models to borrow strength across conditions. There are two main classes of approaches. The penalized methods add various types of penalty terms on the precision matrices of the Gaussian graphical models (GGMs) [Chun, Zhang and Zhao (2015), Danaher, Wang and Witten (2014), Guo et al. (2011)], or in a similar fashion on other types of graphical models [Guo et al. (2015), Xue et al. (2014)], to encourage the commonly shared structures. The Bayesian methods instead incorporate the correlations of different conditions into Markov random field priors [Lin et al. (2017), Mitra, Müller and Ji (2016), Peterson, Stingo and Vannucci (2015)]. Although these methods all allow the graph structures to vary across different conditions, they force all of the samples to share the same graph within each condition. As a result, the heterogeneity of the graph structures within each condition is ignored.

In contrast, there is emerging literature that considers sample heterogeneity within a single condition via clustering approaches. For continuous data, Rodriguez et al. (2011), Cheng and Lenkoski (2012), and Gao et al. (2016) assume that all samples can be generated from several subpopulations, each associated with an unknown GGM. Accordingly, they estimate multiple Gaussian graphs for one condition. For count data, Karlis and Meligkotsidou (2007) develop a finite mixture model of multivariate Poisson distributions. However, they do not investigate the network representation of the dependence structure among variables, and the parameter estimation is impeded by cumbersome computation.

Despite the recent statistical development, rigorous and applicable statistical methods for modeling heterogeneous dynamic TF networks have not yet been established. Previous analysis of ENCODE TF ChIP-seq data attempts to build a single network to represent the co-associations of TFs for each biological condition [Gerstein et al. (2012)]. This approach requires a fixed reference TF and inspects the co-associations of other TFs only within the binding regions of the selected reference TF. Moreover, the resulting networks are not based on a rigorous probabilistic model. For histone modification (HM) ChIP-seq data rather than TF ChIP-seq data, Mitra et al. (2013) build a Markov random field graph to characterize the conditional independence of various types of HMs. Nevertheless, their approach is also constrained to a single condition and ignores the heterogeneity of the networks across genomic locations. Consequently, according to Mitra et al. (2013), their learned graph is highly connected: "In a few cases, however, the strength of the association do not match the hypothesized relationships." These "spurious" associations are likely to be caused by heterogeneity. With the genome classified into several groups, even though the binding events of a TF pair are totally independent within each group, the enforcement of a single network for all genomic locations may induce an edge, indicating the dependence of the two TFs in the network.

Therefore, there is an urgent need to develop statistically rigorous methods to infer heterogeneous dynamic TF networks. The major challenges are summarized as follows. First, ChIP-seq data are count data. Even if we transform the count data into continuous values by adding one and then taking the logarithm, the resulting distribution is right-skewed and heavy-tailed (see the Q–Q plot in Supplementary Material Figure S1 [Luo and Wei (2018)]), thus violating the normality assumption. Therefore, the methods proposed for GGM are not applicable to ChIP-seq count data. Second, due to genome heterogeneity, a single dynamic TF network cannot be assumed for the whole genome. Heterogeneous dynamic TF networks must be estimated across the genome. Third, ChIP-seq experiments consist of very few replicates, and the obtained data are highly noisy. Consequently, to improve signal detection, information should be borrowed across different conditions. However, there is no temporal ordering between biological conditions for the TF association problem. As a result, the relationships among conditions are much more complicated than those for time series data.

To tackle these challenges, we propose a novel Bayesian nonparametric dynamic Poisson graphical model that handles heterogeneity simultaneously across the genome and multiple biological conditions. We design a Markov chain Monte Carlo (MCMC) algorithm to conduct posterior inference and develop a parallel scheme to handle the large data volume. We evaluate the performance of our model via simulation. Finally, we apply the proposed method to ENCODE Tier 1 cell lines to construct TF networks.

2. Model formulation. In this section, we first review the traditional Poisson co-activation graphical model [Xue et al. (2014)], then we generalize it to the dynamic Poisson graphical model for taking multiple conditions into account. Finally, we introduce our hierarchical dynamic Poisson graphical model to deal with multiple conditions and genome heterogeneity at the same time.

2.1. *Poisson co-activation graphical model.* The Poisson co-activation graphical model (PGM) was previously applied successfully to infer the dependence structure of multivariate count data [Karlis (2003), Xue et al. (2014)]. Let $\mathbf{X} = \{X_1, X_2, \ldots, X_p\}$ denote the p-dimensional count data. The dependence of $X_1, X_2, \ldots, X_p$ can be represented by an undirected graph $G = (V, E)$. In graph G, $V = \{1, \ldots, p\}$ is the node set, with node i corresponding to X_i, and $E = (e_{ij})_{1 \le i < j \le p}$ is the edge set, indicating the dependence or independence of X_i and X_j. Specifically, $e_{ij} = 0$ if and only if X_i and X_j are independent. The PGM assumes that $\mathbf{X}$ follows a two-way multivariate Poisson distribution [Karlis (2003)]. Although the general m-way multivariate Poisson is available [Kawamura (1979)] to model higher-order interactions, such as three-way or four-way interactions, as the literature shows, the two-way multivariate Poisson is often sufficient to model the real data reasonably well [Karlis (2003), Karlis and Meligkotsidou (2007), Xue et al. (2014)]. With respect to computation, the time complexity for

estimating a general m-way Poisson distribution is $O(p^m)$ where p is the number of nodes. Hence, the computational time increases dramatically with m, especially when p is large. Therefore, we follow the PGM to focus on two-way multivariate Poisson distributions from this point forward.

For the application of Poisson co-activation graphs to TF networks, p denotes the number of TFs; each node represents a TF; and an edge between nodes i and j indicates that TF i collaborates with TF j when they bind to the DNA.

Recently, there has been active research on the Poisson conditional graphical model (PCGM) [Yang et al. (2013, 2015), Inouye et al. (2017)] as well, which studies whether X_i and X_j are independent conditioning on all the other variables $\mathbf{X} \setminus \{X_i, X_j\}$. Just as both co-expression networks [Carter et al. (2004), Zhang and Horvath (2005), Bickel and Levina (2008)] and Gaussian graphical models [Friedman, Hastie and Tibshirani (2008), Meinshausen and Bühlmann (2006), Yuan and Lin (2007)], measured by covariance and precision matrices, respectively, can characterize different aspects of the gene regulatory mechanism, both PGM and PCGM can offer insights into TF networks. However, for TF cooperation, most existing literature focuses on whether the binding of a given pair of TFs is independent or not without considering other transcription factors [MacArthur et al. (2009), Bickel et al. (2010), Wei and Wu (2016)], which is a close analogy to co-activating brain connectivity networks [Xue et al. (2014)]. Moreover, if only a subset of nodes is measured, the learned partial PCGM network does not reflect the true conditional independence among all nodes. In contrast, the edges in PGM learned on a subset of nodes are the same as the corresponding edges in the complete network learned with data from all of the nodes (see details in Supplementary Material Section 1 [Luo and Wei (2018)]). In the ENCODE project, despite the accumulation of more than 1000 ChIP-seq samples, for each biological condition only a small number of TFs have been assayed: https://genome.ucsc.edu/encode/dataMatrix/encodeChipMatrixHuman.html. In other words, there are many unmeasured nodes in the TF networks. PGM studies whether two TFs are independent or not by themselves, and such interpretations will not change if additional TFs are assayed in the future. Therefore, the currently available data allows a legitimate inference for PGM but not PCGM.

The explicit form of the joint probability function for $\mathbf{X}$ in the PGM is very complicated. For instance, even when $p = 2$, the joint probability for bivariate Poisson distribution [Kocherlakota and Kocherlakota (1992)] is:

$$
\begin{aligned}
&P(X_1 = x_1, X_2 = x_2) \\
&\quad = e^{-(\lambda_{11}+\lambda_{22}+\lambda_{12})} \frac{\lambda_{11}^{x_1}}{x_1!} \frac{\lambda_{22}^{x_2}}{x_2!} \sum_{s=0}^{x_1 \wedge x_2} \binom{x_1}{s} \binom{x_2}{s} s! \left(\frac{\lambda_{12}}{\lambda_{11}\lambda_{22}} \right)^s,
\end{aligned} \tag{2.1}
$$

where $x_1 \wedge x_2$ is the minimum of x_1 and x_2. Fortunately, by data argumentation [Tanner and Wong (1987)], the multivariate Poisson distribution can

be equivalently formulated using a set of independent latent Poisson variables $\mathbf{Y} = \{Y_{ij}\}_{1\le i\le j\le p}$ with corresponding *dependence intensity* parameters $\mathbf{\Lambda} = \{\lambda_{ij}\}_{1\le i\le j\le p}$ [Karlis (2003)]. The observed data are decomposed as $X_i = \sum_{j=1}^{p} Y_{ij}, i = 1, \ldots, p$. Here, $Y_{ij} = Y_{ji}$ for any node pair, representing the symmetry of interaction. Consequently, each X_i marginally follows a Poisson distribution $\text{Pois}(\sum_{j=1}^{p} \lambda_{ij})$. λ_{ij} $(i \neq j)$ characterizes the interaction of TFs i and j, which determines the graph structure. The graph contains an edge $e_{ij} = 1$ if and only if $\lambda_{ij} > 0$. To learn $\mathbf{\Lambda}$ and ultimately the graph G, treating $\mathbf{Y}$ as missing data, the joint probability mass function for the complete data can be written as

$$
\begin{aligned}
f(\mathbf{x}, \mathbf{y}|\mathbf{\Lambda}) &= \Pr(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}|\mathbf{\Lambda}) \\
&= \prod_{i=1}^{p} e^{-\lambda_{ii}} \frac{\lambda_{ii}^{x_i - \sum_{j\neq i} y_{ij}}}{(x_i - \sum_{j\neq i} y_{ij})!} \cdot \prod_{1\le i<j\le p} e^{-\lambda_{ij}} \frac{\lambda_{ij}^{y_{ij}}}{y_{ij}!}.
\end{aligned} \tag{2.2}
$$

An expectation maximization (EM) algorithm [Dempster, Laird and Rubin (1977)] can be derived for the estimation according to the complete log-likelihood. When sparsity is assumed for the graph G, Xue et al. (2014) add a L_1-penalty term $\sum_{1\le i<j\le p} \lambda_{ij}$ to the negative of the complete log-likelihood to shrink the small values in $\mathbf{\Lambda}$ to zero, which enforces a sparse graph structure.

2.2. *Dynamic Poisson graphical model.* For a specific genomic location g, the collaboration pattern of TFs under a given condition d can be characterized by a PGM. Let $X_{dr,i}^{(g)}$ denote the number of ChIP-seq reads for TF i aligned to genomic location g under condition d for replicate r, representing TF i's binding intensity. Without loss of generality, N genomic locations are measured in R replicates for each of the D conditions. In the same fashion as equation (2.2), the joint distribution for count data $\mathbf{X}_{dr}^{(g)} = \{X_{dr,i}^{(g)}\}_{i=1,\ldots,p}$ can be decomposed using latent Poisson random variables $\mathbf{Y}_{dr}^{(g)} = \{Y_{dr,ij}^{(g)}\}_{1\le i\le j\le p}$ as $X_{dr,i}^{(g)} = \sum_{j=1}^{p} Y_{dr,ij}^{(g)}$ for $1 \le i \le p$ with $Y_{dr,ij}^{(g)} = Y_{dr,ji}^{(g)}$. The collaborative efforts of the p TFs are thus characterized by the dependence intensity parameters $\mathbf{\Lambda}_d^{(g)} = \{\lambda_{d,ij}^{(g)}\}_{1\le i\le j\le p}$ associated with $\mathbf{Y}_{dr}^{(g)}$. $\mathbf{\Lambda}_d^{(g)}$ reflects the graph structure $\mathbf{E}_d^{(g)}$ with $\lambda_{d,ij}^{(g)} > 0$ if and only if $e_{d,ij}^{(g)} = 1$.

Because the co-activation patterns of TFs for genomic location g change from one biological condition to another, the condition-specific graph structure $\mathbf{E}_d^{(g)}$ also varies from condition to condition. Collecting $\mathbf{E}_d^{(g)}$ over all of the D conditions into $\mathbf{E}^{(g)} = \{\mathbf{E}_d^{(g)}\}_{d=1,\ldots,D}$, the graph for TFs at genomic location g becomes a *dynamic* one: $G^{(g)} = (V, \mathbf{E}^{(g)})$. Our task is to infer the dynamic graph $\mathbf{E}^{(g)}$ [Figure 1(a)] and the dependence intensity parameters $\mathbf{\Lambda}^{(g)} = \{\mathbf{\Lambda}_d^{(g)}\}_{d=1,\ldots,D}$ from the observed data $\mathbf{X}^{(g)} = \{\mathbf{X}_{dr}^{(g)}\}_{d=1,\ldots,D;r=1,\ldots,R}$ [Figure 1(b)].

According to the PGM, the joint probability mass function for the observed data $\mathbf{X}^{(g)}$ and missing data $\mathbf{Y}^{(g)}$ under multiple conditions becomes

$$
\begin{aligned}
&f\big(\mathbf{x}^{(g)}, \mathbf{y}^{(g)} | \mathbf{\Lambda}^{(g)}, \mathbf{E}^{(g)}\big) \\
&\quad = \prod_{d=1}^{D} \prod_{r=1}^{R} \Bigg[\prod_{i=1}^{p} e^{-\lambda_{d,ii}^{(g)}} \frac{(\lambda_{d,ii}^{(g)})^{(x_{dr,i}^{(g)} - \sum_{j \neq i} y_{dr,ij}^{(g)})}}{(x_{dr,i}^{(g)} - \sum_{j \neq i} y_{dr,ij}^{(g)})!} \\
&\qquad \times \prod_{1 \leq i < j \leq p} e^{-\lambda_{d,ij}^{(g)}} \frac{(\lambda_{d,ij}^{(g)})^{y_{dr,ij}^{(g)}}}{y_{dr,ij}^{(g)}!} \Bigg].
\end{aligned} \tag{2.3}
$$

Although the statistical inference of the dynamic graph can be carried out as before for each single genomic location g separately with the joint probability mass function, the number of replicates of ChIP-seq data for each TF under each condition is very small. The replicate number R is only around two for most ENCODE experiments [ENCODE Project Consortium (2012)]. As a result, analyzing each genomic location individually suffers from low statistical power, and the resulting graph is not only very unstable but is also prone to errors. To improve edge detection for the dynamic graph at each location, the hierarchical model proposed in the next subsection captures the variation of dynamic graphs across the genome and borrows information accordingly.

2.3. *Hierarchical dynamic Poisson graphical model.* The genome is well known to be heterogeneous in terms of GC content, gene density, and HM [Ernst and Kellis (2012)]. Therefore, the co-activation patterns of TFs can be expected to differ among genomic locations. Ignoring the genome heterogeneity and assuming a single dynamic graph for the whole genome, "spurious" associations among TFs will be claimed completely due to heterogeneity as illustrated in Section 4. In the cases where all the genomic locations do share the same dynamic network, our following proposed model will also automatically detect the homogeneity.

To capture the heterogeneity of $\{\mathbf{X}^{(g)} : g = 1, \ldots, N\}$, where N is the number of genomic locations, we impose a Dirichlet process (DP) prior [Ferguson (1973)] on $(\mathbf{\Lambda}^{(g)}, \mathbf{E}^{(g)})$: $P \sim \mathrm{DP}(\alpha H)$; $(\mathbf{\Lambda}^{(g)}, \mathbf{E}^{(g)}) \sim P$ i.i.d. for $g = 1, \ldots, N$. The DP is essentially a probability distribution over a measurable space of probability distributions. The Dirichlet process $\mathrm{DP}(\alpha H)$ is determined by two parameters: α and H. The base distribution H serves as the expected distribution for $(\mathbf{\Lambda}^{(g)}, \mathbf{E}^{(g)})$, and the concentration parameter α controls the deviation from the base distribution. Specifically, as $P \sim \mathrm{DP}(\alpha H)$, for any measurable set A defined on the range space of $(\mathbf{\Lambda}^{(g)}, \mathbf{E}^{(g)})$, the probability $P(A)$ of set A as a random variable follows a beta distribution $\mathrm{Beta}(\alpha H(A), \alpha[1 - H(A)])$ with expectation $H(A)$. The DP prior for the dependence intensity parameters in dynamic PGMs helps to build a *hierarchical dynamic Poisson graphical model* (HDPGM) to characterize the variations and

similarities of TF collaborations across the genome. The general version of the HDPGM is formulated as follows:

$$P \sim \mathrm{DP}(\alpha H);$$
$$\left(\mathbf{\Lambda}^{(g)}, \mathbf{E}^{(g)}\right) \sim P \qquad \text{i.i.d.};$$
$$\left(\mathbf{X}^{(g)}, \mathbf{Y}^{(g)}\right) \sim f\left(\mathbf{x}^{(g)}, \mathbf{y}^{(g)} | \mathbf{\Lambda}^{(g)}, \mathbf{E}^{(g)}\right),$$

where f is specified in Equation (2.3).

The Chinese restaurant process metaphor [Aldous (1985)] for DP naturally offers a procedure for clustering genomic locations. Accordingly, the HDPGM can be described by the following data-generating process. There are infinitely many tables in a Chinese restaurant, and table k is associated with parameters $\mathbf{\Theta}_k^* = \{\mathbf{\Theta}_{kd}^*\}_{d=1,\ldots,D}$ sampled from the base distribution H, where $\mathbf{\Theta}_{kd}^* = \{\theta_{kd,ij}, L_{kd,ij}\}_{1\le i\le j\le p}$. $L_{kd,ij}$ $(i < j)$ indicates whether there is an edge between nodes i and j under condition d for table k. When $L_{kd,ij} = 0$, $\theta_{kd,ij} = 0$. When $L_{kd,ij} = 1$, $\theta_{kd,ij}$ is positive and describes the association degree between nodes i and j under condition d for table k. Genomic location one randomly chooses a table and sits there. The genomic location g $(g \ge 2)$ is assigned to an occupied table with a probability proportional to the number of previous genomic locations sitting at that table and allocated to a new table with a probability proportional to α. After obtaining its table label $C^{(g)}$, genomic location g generates its latent variables $\mathbf{Y}_{dr}^{(g)}$ for condition d according to a multivariate Poisson with parameters $\mathbf{\Theta}_{C^{(g)}d}^*$, respectively. In other words, for $1 \le d \le D$, $(\mathbf{\Lambda}_d^{(g)}, \mathbf{E}_d^{(g)}) = \mathbf{\Theta}_{C^{(g)}d}^* = (\mathbf{\Theta}_{C^{(g)}d}, \mathbf{L}_{C^{(g)}d})$ in equation (2.3). Finally, the observed data $\mathbf{X}_{dr}^{(g)}$ come from the latent $\mathbf{Y}_{dr}^{(g)}$. On the one hand, genomic locations that belong to the same table share the same multivariate Poisson distributions and dynamic graph structure. On the other hand, the different tables capture the heterogeneity.

By accounting for network heterogeneity, the HDPGM inherently models over-dispersion. Over-dispersion is a major concern when the sequencing data are fitted with a single Poisson distribution. As a result, the negative binomial is often used instead [Robinson, McCarthy and Smyth (2010)]. A negative binomial can be viewed as a continuous mixture of Poisson distributions whose mixing distribution of the Poisson rate is a gamma distribution. In comparison, in the HDPGM, the marginal distribution of $\mathbf{X}_{dr}^{(g)}$ is a Poisson mixture in which the mixing distribution is nonparametric and hence more flexible. Therefore, the HDPGM naturally handles over-dispersion.

Moreover, although the traditional PGM has been criticized for allowing only positive dependencies [Inouye et al. (2017)], HDPGM is able to model both positive and negative dependencies for count data. The heterogeneity considered by HDPGM helps to relax the strictly positive correlation between variables imposed by PGM. We illustrate this concept with a simple example in the Supplementary Material Section 2 [Luo and Wei (2018)].

In addition to tackling over-dispersion and positive correlation restriction, our model automatically borrows information across both biological conditions and genomic locations to improve the signal detection for dependence intensity parameters. Let us look at the joint distribution of latent variables $Y^{(g)}_{cr,ij}$ and $Y^{(g)}_{dr',ij}$ which correspond to the interaction strength between nodes i and j at genomic location g under conditions c and d, respectively. The stick-breaking representation [Ishwaran and James (2001)] of DP allows us to construct $P \sim \mathrm{DP}(\alpha H)$ as follows: $P(\cdot) = \sum_{k=1}^{\infty} \pi_k \delta_{\mathbf{\Theta}^*_k}(\cdot)$, $\pi_k = V_k \prod_{l<k}(1 - V_l)$, $V_k \sim \mathrm{Beta}(1, \alpha)$, $\mathbf{\Theta}^*_k \sim H$, where $\delta_a(\cdot)$ is the Dirac distribution at a. Conditioning on $\mathbf{V} = \{V_1, V_2, \ldots\}$, and hence equivalently $\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots\}$, we have $p(Y^{(g)}_{cr,ij} = a, Y^{(g)}_{dr',ij} = b|\boldsymbol{\pi}, \mathbf{\Theta}^*) = \sum_k [\pi_k e^{-\theta_{kc,ij}} \frac{(\theta_{kc,ij})^a}{a!} e^{-\theta_{kd,ij}} \frac{(\theta_{kd,ij})^b}{b!}]$. However, on the other hand, $p(Y^{(g)}_{cr,ij} = a|\boldsymbol{\pi}, \mathbf{\Theta}^*) \cdot p(Y^{(g)}_{dr',ij} = b|\boldsymbol{\pi}, \mathbf{\Theta}^*) = [\sum_k \pi_k e^{-\theta_{kc,ij}} \frac{(\theta_{kc,ij})^a}{a!}][\sum_k \pi_k e^{-\theta_{kd,ij}} \frac{(\theta_{kd,ij})^b}{b!}] \neq p(Y^{(g)}_{cr,ij} = a, Y^{(g)}_{dr',ij} = b|\boldsymbol{\pi}, \mathbf{\Theta}^*)$. Therefore, $Y^{(g)}_{cr,ij}$ and $Y^{(g)}_{dr',ij}$ are no longer independent under the HDPGM, and strong signals under one condition improve the edge detection under the other. For instance, suppose the network structures of clusters 1 and 2 are very similar under condition 1 but more distinct under condition 2. Compared to the separate analysis under condition 1, the joint analysis offers better clustering accuracy by using data from condition 2, which in turn helps to detect the subtle edge difference between clusters 1 and 2 under condition 1. Thus, the joint analysis of D conditions differs from analyzing each condition separately and improves edge detection. In the simulation study, we demonstrate the advantages of the joint analysis with an even more difficult scenario where the network structures between two clusters are always close across all the conditions. Meanwhile, according to the DP, genomic locations assigned to the same table are governed by the same $\mathbf{\Theta}^*_k$. Thus, information is also shared across genomic locations. Because of the two-way information pooling, the proposed model is able to learn the underlying dynamic graph from the highly noisy ChIP-seq data.

3. Statistical inference. In this section, we conduct statistical inference for our model HDPGM in the full Bayesian framework.

3.1. *Prior specification.* We now specify the base distribution H for the Dirichlet Process $\mathrm{DP}(\alpha H)$. Recall that in the Chinese restaurant process table k is associated with a dynamic TF network structure $\mathbf{\Theta}^*_k = \{\mathbf{\Theta}^*_{kd}\}_{d=1,\ldots,D}$, which is sampled from H. Here, we assume independent priors for parameters related to each edge and node:

$$H(\mathbf{\Theta}^*_{kd}) = \bigotimes_{i=1}^{p} H_{ii}(\theta_{kd,ii}) \bigotimes_{1\le i<j\le p} H_{ij}(\theta_{kd,ij}, L_{kd,ij}) \qquad \text{for } 1 \le d \le D. \tag{3.1}$$

For each node, we assume $H_{ii}(\theta_{kd,ii})$ is a gamma distribution $\Gamma(\tau_2, \gamma_2)$. For each edge, we assign a Bernoulli prior $\mathrm{Ber}(q)$ to the edge indicator $L_{kd,ij}$. If $L_{kd,ij} = 0$,

$\theta_{kd,ij}$ is assumed to follow a gamma distribution $\Gamma(\tau_0, \gamma_0)$ with a mean close to zero and a small variance; if $L_{kd,ij} = 1$, then $\theta_{kd,ij}$ comes from a gamma distribution $\Gamma(\tau_1, \gamma_1)$ with a relatively large mean and a large variance. In this way, the base distribution H assumes that $\theta_{kd,ij}$ $(i \neq j)$ marginally follows a spike-slab prior [George and McCulloch (1993), Ishwaran and Rao (2005)] in the form of a gamma mixture $q\Gamma(\tau_1, \gamma_1) + (1-q)\Gamma(\tau_0, \gamma_0)$.

3.2. *Posterior computation*. To explore the posterior distribution, we propose a hybrid MCMC algorithm, built on the blocked Gibbs sampler developed for the truncation approximation to the DP [Ishwaran and James (2001)]. Specifically, based on the stick-breaking representation, the DP can be approximated accurately by setting $V_M = 1$ for a sufficiently large M so that $\pi_k = 0$ for $k \geq M+1$.

We augment a random variable $C^{(g)}$ for each genomic location, indicating its table label as described in the Chinese restaurant process metaphor. In other words, $C^{(g)} = k$ if and only if $(\mathbf{\Lambda}^{(g)}, \mathbf{E}^{(g)}) = \mathbf{\Theta}^*_k$. Collectively, $\mathbf{C} = \{C^{(g)}\}_{g=1,\ldots,N}$ record the class memberships for all genomic locations. We can derive the following blocked hybrid Gibbs sampler at iteration t:

1. Update the dynamic TF network structure $\mathbf{\Theta}^*_k$ for $k = 1, \ldots, M$: let $\mathbf{C}^{*[t-1]}$ be the unique set of $\mathbf{C}^{[t-1]}$. If $k \notin \mathbf{C}^{*[t-1]}$, sample $\mathbf{\Theta}^{*[t]}_k$ from the base distribution H of the DP. If $k \in \mathbf{C}^{*[t-1]}$, sample $\mathbf{\Theta}^{*[t]}_k \propto H(\mathbf{\Theta}^{*[t]}_k) \prod_{\{g:C^{(g),[t-1]}=k\}} f(\mathbf{x}^{(g)}, \mathbf{y}^{(g),[t-1]} | \mathbf{\Theta}^{*[t]}_k)$, where f follows equation (2.3) and H is specified as in equation (3.1).
2. Update the class membership $C^{(g)}$ for $g = 1, \ldots, N$.
3. Update the class proportion $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)$: sample the stick-breaking weights $V^*_k \sim \text{Beta}(\frac{\alpha}{M} + M^{[t]}_k, \frac{M-k}{M}\alpha + \sum_{l=k+1}^{M} M^{[t]}_l)$ for $k = 1, \ldots, M-1$. $M^{[t]}_l = \#\{g : C^{(g),[t]} = l\}$ is the number of genomic locations assigned to table l in iteration t. Once all of the V_k, $k = 1, \ldots, M-1$, are sampled, the cluster abundance $\boldsymbol{\pi}^{[t]}$ is updated as $\pi^{[t]}_1 = V^*_1$, $\pi^{[t]}_k = (1 - V^*_1)\ldots(1 - V^*_{k-1})V^*_k$ for $k = 2, \ldots, M-1$, and $\pi^{[t]}_M = 1 - \sum_{k=1}^{M-1} \pi^{[t]}_k$.
4. Update the underlying latent counts $\mathbf{Y}$: sample matrix $\mathbf{Y}^{(g),[t]}_{dr}$ from $f(\mathbf{x}^{(g)}_{dr}, \mathbf{Y}^{(g),[t]}_{dr} | \mathbf{\Theta}^{*[t]}_{C^{(g),[t]}})$ based on a Metropolis–Hastings step.

Sampling $\mathbf{\Theta}^* = (\mathbf{\Theta}, \mathbf{L})$, $\mathbf{C}$, and $\boldsymbol{\pi}$ from exponential families is straightforward. However, the conditional functions for $\mathbf{Y}$ do not belong to any standard distribution, so we incorporate a Metropolis–Hastings step [Metropolis et al. (1953)] with a computational complexity of $O(DRp^2)$ for each genomic location. Therefore, the overall computational complexity for one Gibbs cycle is $O(MNDRp^2)$. The details of the MCMC algorithm are listed in the Supplementary Material Section 3 [Luo and Wei (2018)].

3.3. *Parallel computing*. The large volume of data requires intensive computation and a large amount of memory. Therefore, the fast computation for the posterior calls for parallelization. Fortunately, unlike general MCMC, which is constrained to sequential operations, our blocked hybrid Gibbs sampler allows efficient parallel programming.

We develop a parallel programming algorithm under the message-passing framework [Gropp, Lusk and Skjellum (1999)]. In our system, there are a "master" processor and S "worker" processors. The whole genome is divided into S portions, and all of the $\mathbf{X}^{(g)}$s within the same portion and their corresponding $\mathbf{Y}^{(g)}$s and $C^{(g)}$s are assigned to a single worker processor. In contrast, $\mathbf{V} = \{V_k\}_{k=1,\ldots,M-1}$ and $\mathbf{\Theta}^* = \{\mathbf{\Theta}_k^*\}_{k=1,\ldots,M}$ are stored on both the master processor and all of the S worker processors. For each genomic location, Steps 2 and 4 of the sampler rely only on local $\mathbf{X}^{(g)}$ and $\mathbf{Y}^{(g)}$ information as well as the copies of $\mathbf{V}$ and $\mathbf{\Theta}^*$ on its own assigned worker. Therefore, Steps 2 and 4 can be conducted simultaneously across all genomic locations on all of the S worker processors. Steps 1 and 3 involve updating global parameters for the model. For the proposed model, summary statistics can be computed at each worker first and then sent to the master via message-passing. For instance, in Step 3, updating V_k asks for sampling from $\text{Beta}(\frac{\alpha}{M} + M_k, \frac{M-k}{M}\alpha + \sum_{l=k+1}^{M} M_l)$. To count M_l, on each processor s, the number of genomic locations assigned to table l can be counted as $M_{l,s}$. After the master processor receives all $M_{l,s}, s = 1, \ldots, S$, M_l can be computed as $M_l = \sum_{s=1}^{S} M_{l,s}$. Next, a new V_k is sampled at the master and broadcasted back to each worker processor. Each worker processor updates its local copy of V_k after receiving the message from the master processor. Step 1 can be implemented in the same way. The proposed algorithm is highly scalable.

3.4. *Graph inference*. We make posterior inferences based on the samples from the MCMC algorithm. We assume that there are a total of T iterations and that the samples in the last B iterations are collected after a burn-in period. Note that the DP allows the number of occupied tables, or cluster number K, to be random. From the posterior samples, we take the posterior mode $\hat{K}$ as the estimated number of dynamic networks presented in the data.

For edge selection, conditional on $\hat{K}$ and given an occupied table k, we calculate the posterior marginal probability of including edge $e_{ij} (i < j)$, denoted by $\text{PPI}_{kd,ij}$ [Peterson, Stingo and Vannucci (2015)], using the posterior samples of $L_{kd,ij}^{[T-t+1]}$ $(t = B, B-1, \ldots, 1)$. Specifically, $\text{PPI}_{kd,ij} = P(L_{kd,ij} = 1|\mathbf{X}) \approx \frac{1}{B}\sum_{t=1}^{B} L_{kd,ij}^{[T-t+1]}$. Accordingly, following Peterson, Stingo and Vannucci (2015), the expected false discovery rate (FDR) for edge detection can be estimated as

$$\text{FDR}(\kappa) = \frac{\sum_{d=1}^{D}\sum_{k=1}^{K}\sum_{i<j}\xi_{kd,ij} I(\xi_{kd,ij} \le \kappa)}{\sum_{d=1}^{D}\sum_{k=1}^{K}\sum_{i<j} I(\xi_{kd,ij} \le \kappa)}, \tag{3.2}$$

where $\xi_{kd,ij} = 1 - \text{PPI}_{kd,ij}$ [Newton et al. (2004)], κ is the cutoff value to call an edge, and $I(\cdot)$ is the indicator function. We select κ such that $\text{FDR}(\kappa) \leq \alpha_0$, where α_0 is a small value, such as 0.01. If $\xi_{kd,ij} \leq \kappa$, there is an edge between nodes i and j in the cluster k under condition d; otherwise, there is no edge. In other words, when $\xi_{kd,ij} \leq \kappa$, we believe that TF i collaborates with TF j in the TF network k under biological condition d; otherwise, there is no co-activation between them. When an edge is learned, we also estimate its intensity parameter $\theta_{kd,ij}$ with the posterior mean $\hat{\theta}_{kd,ij}$; otherwise, $\theta_{kd,ij}$ is estimated to be zero. Throughout the paper, we follow the tradition of taking $\kappa = 0.5$ as default [Peterson, Stingo and Vannucci (2015)] unless it fails to control the estimated FDR in equation 3.2 below the target threshold α_0. In the latter case, we state κ explicitly. For clustering, genomic location g is assigned according to the posterior mode of $C^{(g),[t]}$s.

4. Simulation. We conduct a simulation study to evaluate the performance of HDPGM in correct detection of heterogeneous subpopulations, accurate estimation of model parameters, and precise recovery of the underlying TF network structures.

We simulate $N = 5000$ genomic locations coming from $K = 4$ subpopulations, that is, four clusters. A genomic location is assigned to cluster 1 with a probability of 30%, to cluster 2 with 20%, to cluster 3 with 25%, and to cluster 4 with 25%. For each genomic location, we simulate read counts for $p = 10$ TFs under $D = 3$ conditions with $R = 2$ replicates. In other words, the TF network on each genomic location consists of ten nodes and varies across three conditions. The TF network structures $\mathbf{L}_{kd}$ and underlying true dependence intensity parameters $\boldsymbol{\Theta}_{kd}, d = 1, \ldots, D, k = 1, \ldots, K$ are presented in Figure 2(a) and Figure 3(a), respectively. Note that the difference between clusters 1 and 2 is slight, and the differential edges are highlighted in red in Figure 2(a). The detailed steps for generating the simulation data are listed in the Supplementary Material Section 4 [Luo and Wei (2018)].

We set the truncation parameter for DP as $M = 10$ and run the MCMC algorithm for 100,000 iterations, which took 4.72 hours for parallel computing with 10 cores. The hyper-parameters $(\tau_0, \gamma_0, \tau_1, \gamma_1, \tau_2, \gamma_2)$ are taken as $(2, 20, 2, 1, 3, 1)$, and q is set as 0.25. We discard the first 50,000 iterations as burn-in. The trace plots indicate that our algorithm has a good convergence property (see Supplementary Material Figure S4 (a–b) [Luo and Wei (2018)]). From the collected samples, the mode for cluster number K (the number of occupied tables in the Chinese Restaurant metaphor) is four. Thus, our algorithm correctly identifies the number of heterogeneous subpopulations in the data. After accounting for label switching, we then focus on all MCMC samples for which $K = 4$. For simulation data, as we know the underlying true status of each edge, we can calculate the true FDR by counting the number of false positives among all claimed edges, which turns out to be zero. The corresponding intensity parameters are estimated accordingly.

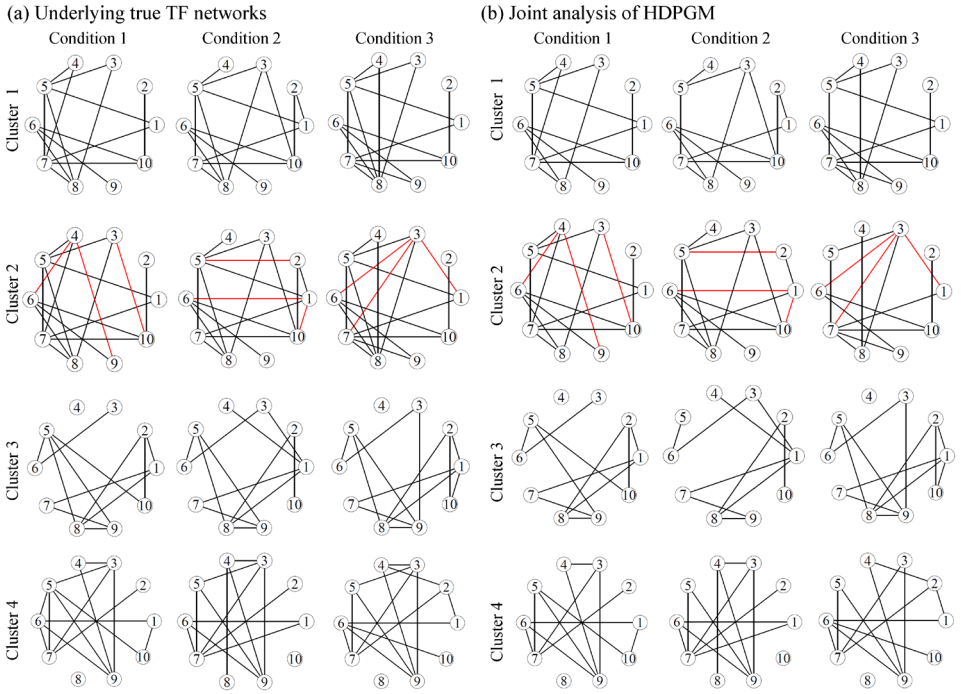


FIG. 2. *The underlying true and estimated TF network structures in the simulation study.* (a) *The underlying true TF networks. TF networks in cluster* 1 *are similar to TF networks in cluster* 2, *and the differential edges between cluster* 1 *and cluster* 2 *are highlighted in red.* (b) *The estimated TF networks by the joint analysis. The graph in cluster k and condition d in panel* (b) *is an estimation of the graph in cluster k and condition d in panel* (a). *Although the differential signals between cluster* 1 *and cluster* 2 *are slight under each condition, borrowing information across all of the three conditions can detect the differential edges which are colored in red.*

The estimated TF network structures and associated dependence intensity parameters across multiple conditions are demonstrated in Figure 2(b) and Figure 3(b), respectively. From Figures 2 and 3, we can see that the underlying graph structures are well recovered. Moreover, 98.6% of the samples are correctly grouped based on class membership indicator estimates $\hat{C}^{(g)}$s. To quantify the precision of the estimation for $\boldsymbol{\Theta}_{kd}$, we simulate the dataset five times and then obtain five sets of $\boldsymbol{\Theta}_{kd}$ estimations. Subsequently, the root mean square errors (RMSEs) for each dependence intensity parameter are calculated (see Supplementary Material Tables S1–S12 [Luo and Wei (2018)]). Most of the RMSEs are small, indicating that HDPGM performs well on the estimation of dependence intensity parameters.

We compare HDPGM with three traditional approaches to estimating TF networks: (I) learning a TF network for each genomic location under each condition separately, (II) learning a single TF network for all genomic locations given an individual condition, and (III) learning multiple heterogeneous TF networks for an individual condition without pooling information across conditions. For Type

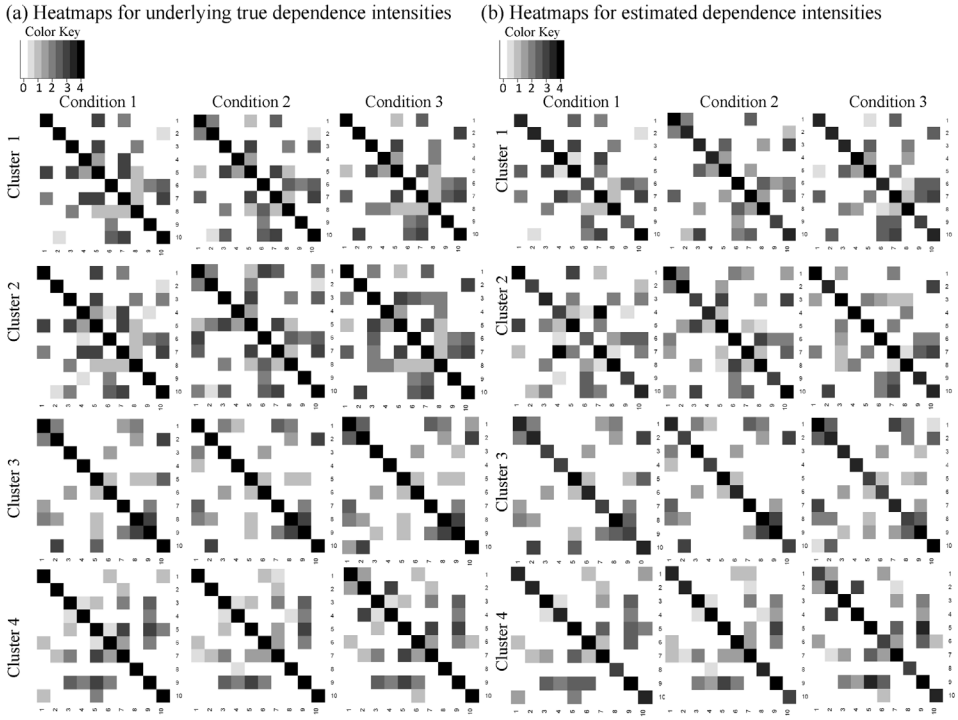


FIG. 3. *The underlying true and estimated dependence intensities in the simulation study.* (a) *The underlying true dependence intensities*. (b) *The estimated dependence intensities*. *The heatmap in cluster k and condition d in panel* (b) *is an estimation of the truth in cluster k and condition d in panel* (a).

I analysis, we run the MATLAB program of the EM algorithm provided by Xue et al. (2014) to each genomic location under each condition. As a result, we obtain estimations of dependence intensity parameters for 5000(genomic locations) $*$ 3(conditions) TF networks. For Type II analysis, we learn a single TF network for each condition with the same MATLAB program [Xue et al. (2014)] again, which results in parameter estimations for 3(conditions) TF networks. To verify the importance of borrowing strengths across multiple conditions ($D > 1$) in HDPGM, we carry out Type III analysis by separately applying our HDPGM algorithm to the data under each condition.

For both Type I and Type II analyses, given a cutoff δ, we claim an edge exists between node i and j if its corresponding dependence intensity parameter estimate is greater than δ. By varying δ, we obtain receiver operating characteristic (ROC) curves [Hanley and McNeil (1982)] for the two methods, respectively. For Type III analysis and HDPGM, since an edge is called according to κ in equation (3.2), we can also draw the ROC curves by varying κ. Figure 4(a) shows that the ROC curve for HDPGM is above the other three ROC curves. Type I analysis suffers from the

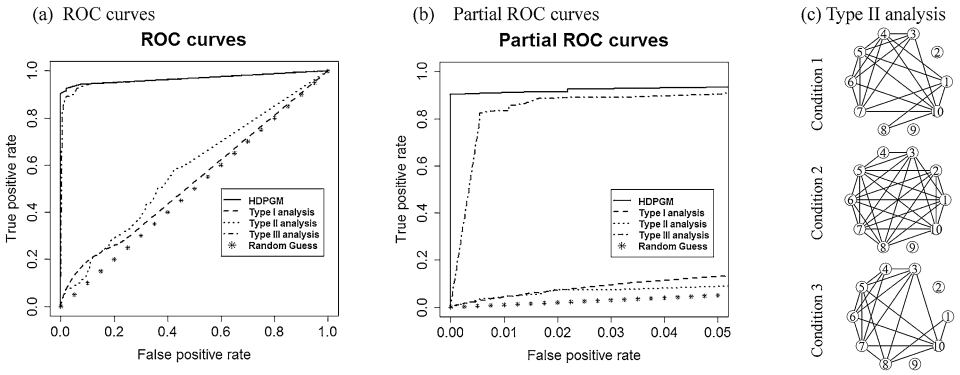


FIG. 4. (a) *ROC curves in terms of edge detection for HDPGM and three traditional types of analyses*; (b) *Partial ROC curves by controlling FPR less than* 0.05; (c) *The three TF networks learned by Type II analysis where the cutoff* δ *to call an edge is* 0.01.

very low replicate numbers for a single genomic location under a single condition. In contrast, Type II analysis pools information across all genomic locations within a condition, but it ignores the inherent heterogeneity among genomic locations, leading to many spurious associations [see Figure 4(c)]. For Type III analysis, as the assumed network structures are similar for clusters 1 and 2 across the three conditions [see Figure 2(a)], it becomes hard for the single condition-based analysis to distinguish between the two clusters and to detect differential weak signal edges between clusters 1 and 2 as shown in Supplementary Material Table S13 [Luo and Wei (2018)]. In comparison, when applying HDPGM to jointly model multiple conditions, the misclassification rate is only 1.4% and the joint analysis successfully detects all of the differential edges between clusters 1 and 2 as shown in red in Figure 2(b). We further examine the ROC curves where the false positive rate (FPR) is controlled less than 0.05 [Figure 4(b)], which is of prime interest for edge detection. Figure 4(b) shows that at low FPRs HDPGM can achieve much larger statistical power than the other three types of analyses. Hence, borrowing strengths across conditions and genomic locations improves edge detection substantially.

The simulation study illustrates that the HDPGM is able to discover heterogeneous subpopulations and borrow information across conditions and genomic locations to improve edge detection. In particular, it highlights the danger of ignoring heterogeneity across the genome and emphasizes the urgency of adopting the HDPGM to examine TF associations.

4.1. *Sensitivity analysis*. In this subsection, we investigate how different choices of hyper-parameters affect the performance of HDPGM. There are seven hyper-parameters in HDPGM: $(q, \tau_0, \gamma_0, \tau_1, \gamma_1, \tau_2, \gamma_2)$. q is the prior probability that an edge is present in the TF network. (τ_0, γ_0), (τ_1, γ_1), and (τ_2, γ_2) correspond

to the "(shape, rate)" parameters in the prior gamma distributions. Accordingly, $\frac{\tau_0}{\gamma_0}$ and $\frac{\tau_1}{\gamma_1}$ are the prior means of $\theta_{kd,ij}$ $(i \neq j)$ for $L_{kd,ij} = 0$ and $L_{kd,ij} = 1$, respectively. When $L_{kd,ij} = 0$, $\theta_{kd,ij}$ should be small enough and well separated from $\theta_{kd,ij}$ when $L_{kd,ij} = 1$. Following a similar approach in [George and McCulloch (1993)], we let the prior mean ratio $\frac{\tau_0/\gamma_0}{\tau_1/\gamma_1}$ be c. The hyper-parameter c partially describes how far $\Gamma(\tau_0, \gamma_0)$ is away from $\Gamma(\tau_1, \gamma_1)$. The smaller c, the closer $\Gamma(\tau_0, \gamma_0)$ is to zero compared to $\Gamma(\tau_1, \gamma_1)$. Taking this into account, we let c be between 0 and 0.5. We set $\tau_0 = \tau_1 = 2$ to guarantee the existence of the modes of gamma distributions and set γ_0 to 20 such that $\frac{\tau_0}{\gamma_0} = 0.1$, close to zero. Consequently, $(q, \tau_0, \gamma_0, \tau_1, \gamma_1, \tau_2, \gamma_2) = (q, 2, 20, 2, 20c, \tau_2, \gamma_2)$. Compared to q and c, the choices of τ_2 and γ_2 are not so important in terms of network recovery because they correspond to the diagonals in the TF networks. Therefore, we set $(\tau_2, \gamma_2) = (3, 1)$. Finally, we investigate the influence of the different choices of q and c.

We let the hyper-parameter q vary from 0.1 to 0.9 with step 0.1, and let c change from 0.1 to 0.5 with step 0.1, so we have a total of 45 choices for (q, c). Given a set of (q, c), we then run the HDPGM algorithm. We compare the different choices of (q, c)s in clustering (ARI) and network recovery (TPR, FPR, and FDR) accuracy. Supplementary Material Figure S5(a) [Luo and Wei (2018)] shows that the clustering result was robust to the different choices of (q, c). For edge detection, on the one hand, the large q and c lead to a large FPR or FDR (see Supplementary Material Figure S5(c–d) [Luo and Wei (2018)]); on the other hand, when q is between 0.1 and 0.4, TPR is not very sensitive to the choice of c (see Supplementary Material Figure S5(b) [Luo and Wei (2018)]). Therefore, to achieve reasonable performance, we recommend setting the hyper-parameter q to be smaller than 0.5 and the ratio c to be less than 0.3.

4.2. *Model misspecification.* Now we investigate how HDPGM performs when the model is misspecified. First, we consider a scenario where the count data are generated from three-way multivariate Poisson distributions. We let $X_i^{(g)} = \sum_{j=1}^{p} Y_{ij}^{(g)} + \sum_{j,k=1}^{p} Y_{ijk}^{(g)}$ in this case rather than $X_i^{(g)} = \sum_{j=1}^{p} Y_{ij}^{(g)}$ as in the two-way multivariate Poisson distribution. $Y_{ijk}^{(g)}$ represents the three-way interaction for TFs i, j, and k. We introduce δ to denote the number of three-way interactions present in the model-misspecified dataset. From Supplementary Material Figure S6 [Luo and Wei (2018)], we can see that as long as there are not too many three-way interactions and their intensities are moderate, HDPGM can still provide proper clustering accuracy.

We further investigate another misspecified setting where there exist spatial correlations between nearby genomic locations (and each genomic location follows a two-way multivariate Poisson). We designed two datasets. In the first dataset, we encourage the nearby genomic locations to come from the same cluster with high

probability. Moreover, we let the observed count data of nearby locations be correlated (see Supplementary Material Section 5 [Luo and Wei (2018)] for details on the data generation procedure). In the second dataset, the cluster memberships for all of the locations are drawn independently. Given the cluster indicators, nearby genomic locations have independent counts. We then applied the proposed model HDPGM to both datasets. It turns out that not considering the spatial correlations in HDPGM can lose some power when inferring the network structure (see Supplementary Material Figure S7 [Luo and Wei (2018)]). However, the power loss is not substantial, and HDPGM can still give a good clustering result.

5. Application. By August 30, 2017, a total of 30 TFs in the ENCODE project are assayed in all of the three Tier 1 cell lines: GM12878, H1-hESC, and K562 (see ENCODE ChIP-seq data matrix: https://genome.ucsc.edu/encode/dataMatrix/encodeChipMatrixHuman.html). Thus, we apply the HDPGM to study the collaborations of the 30 TFs across the three cell lines. Each TF under each condition has two replicates. We download $p \times D \times R = 30 \times 3 \times 2 = 180$ aligned BAM files from the ENCODE project. Here we are specifically interested in the TF collaborations at promoter regions. Therefore, we take the aforementioned genomic locations as the 1000 base-pair-long bins upstream of the transcription start sites and count the number of reads aligned to each genomic location (i.e., promoter region). Consequently, we obtain a ChIP-seq count table as illustrated in Figure 1(b) for 30 TFs at a total of 22,402 genomic locations. Details of the data preprocessing are presented in Supplementary Material Section 6 [Luo and Wei (2018)].

When applying HDPGM, we set the truncation number $M = 40$, the hyper-parameters in the gamma distributions $(\tau_0, \gamma_0, \tau_1, \gamma_1, \tau_2, \gamma_2) = (2, 20, 2, 1, 3, 1)$, and the hyper-parameters in the Bernoulli distribution $q = 0.25$. Sensitivity analysis (see Supplementary Material Figure S8 [Luo and Wei (2018)]) shows that the clustering results are robust to q when q is between 0.05 and 0.45. After 50,000 burn-in iterations, we collected another 50,000 samples for posterior inference. In total, the HDPGM algorithm took 4.11 days with parallel computing using 15 cores. The trace plots are shown in Supplementary Figure S9 [Luo and Wei (2018)]. The HDPGM discovers 37 clusters of dynamic TF networks for all of the promoter regions. We let $\kappa = 0.08$ such that the estimated FDR is 0.0092 based on equation 3.2. Therefore, few claimed edges are false positives. Four representative dynamic TF networks are shown in Figure 5.

Figure 5 confirms that TF collaborations vary across conditions. In cluster 3, MXI1 interacts with MAX and MAX collaborates with MYC in cell lines GM12878 and H1-hESC, but the interaction between MXI1 and MAX is broken in K562. It is noteworthy that K562 is a cancer cell line while GM12878 and H1-hESC are normal cell lines. Meanwhile, it is known that MXI1-MAX heterodimers hinder the function of MYC [Zervos, Gyuris and Brent (1993)], and MYC is associated with various human cancers [Grandori et al. (2000)]. Consequently, our

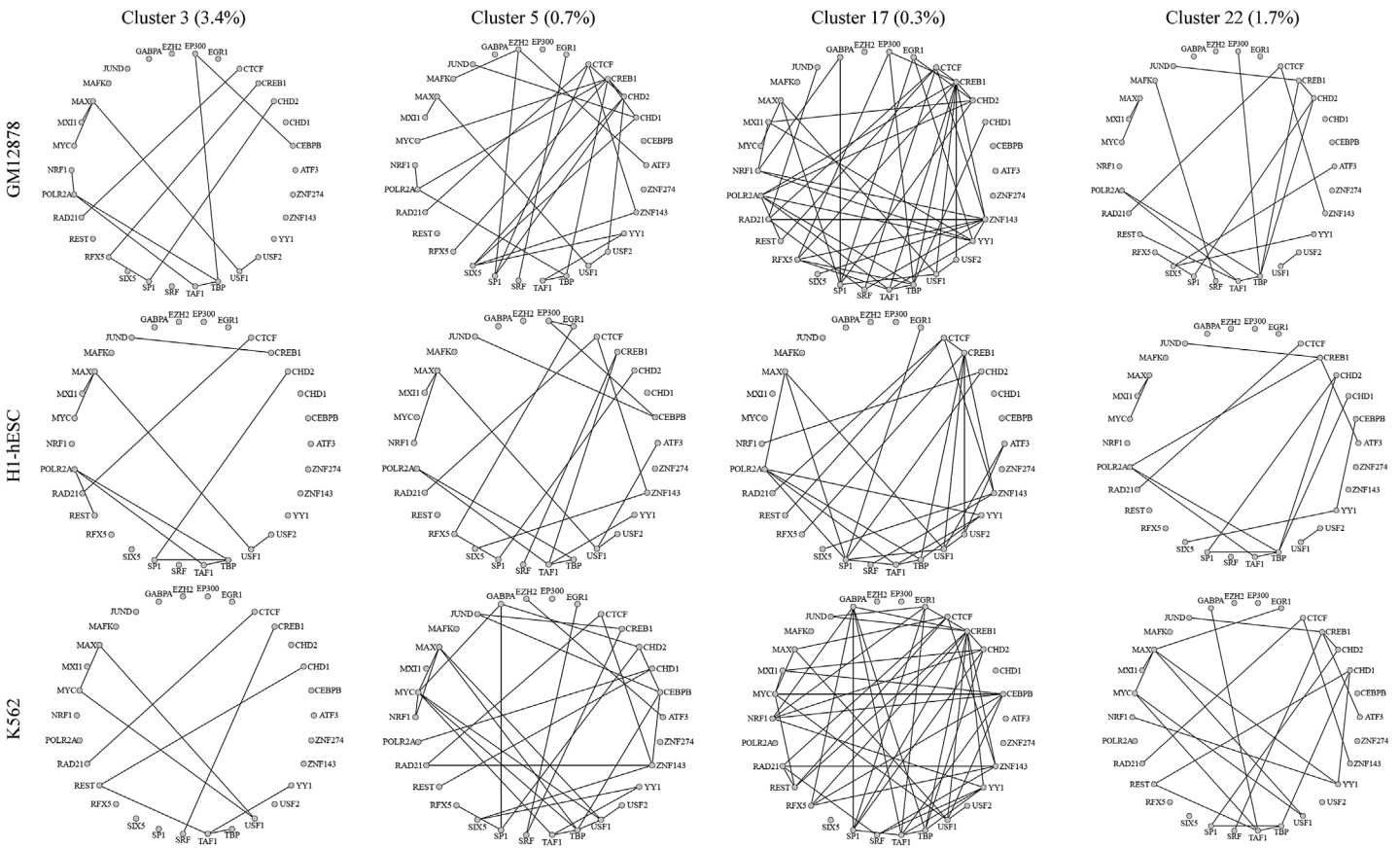


FIG. 5. *Four representative TF networks from three types of Tier* 1 *cell lines*: *GM*12878, *H*1-*hESC*, *and K*562. *The number in the parentheses indicates the estimated proportion of the corresponding cluster.*

analyses suggest that the carcinogenesis of K562 may be related to the broken collaboration between MAX and MXI1, which activates MYC and triggers the expression of its target oncogenes.

Moreover, the TF collaborations are also heterogeneous across the genome. For cell line K562, MAX and CTCF strongly co-activate at the promoter regions in cluster 17, but they bind to the genome independently at promoters in clusters 3, 5, and 22. Meanwhile, MYC and REST interact with each other in cluster 17 but not in clusters 3, 5, or 22. Similarly, in H1-hESC and K562, CREB1 does not collaborate with RFX5 in clusters 3, 9, and 22, but they interact in cluster 17. A previous study discovered that the interaction of CREB and RFX5 regulates the expression of MHC class II genes [Lochamy, Rogers and Boss (2007)]. In cluster 17, we find the promoter regions of 12 out of the total 15 MHC class II genes, implying a significant enrichment of cluster 17 for MHC class II genes (p-value $= 2.2 \times 10^{-16}$ with Fisher's exact test) and confirming the findings of Lochamy, Rogers and Boss (2007). These phenomena justify that the heterogeneity of TF networks does exist among genomic locations.

We further conduct the pathway enrichment analysis [Subramanian et al. (2005)] for genes in each cluster on http://software.broadinstitute.org/gsea/msigdb/annotate.jsp using KEGG gene sets [Ogata et al. (1999)]. The enriched pathways for each cluster are listed in Supplementary Material Tables S14–S22 [Luo and Wei (2018)]. One interesting discovery is that there are 10 clusters (including cluster 3 and cluster 22) enriched with the KEGG HUNTINGTONS DISEASE pathway. SP1 and TBP are two of the main TFs that can interact with the huntingtin protein, and their abnormal interactions may lead to Huntington's disease [Li and Li (2004)]. In the networks of the 10 clusters associated with the KEGG HUNTINGTONS DISEASE pathway, TBP firmly connects to TAF1 across all networks, and SP1 associates with CHD2 in most of the networks.

To evaluate how well HDPGM fits the data, we compare the marginal distribution of $X_{dr,i}$ estimated by the HDPGM to the observed distribution for each TF i under each condition d. The marginal distribution of $X_{dr,i}$ is generated as follows. First, we estimate the dependence intensity parameters $\hat{\theta}_{kd,ij}$ and the cluster indicators $\hat{C}^{(g)}$. Next, for each genomic location g, we sample pseudo-data $X^{*(g)}_{dr,i}$ from $\text{Pois}(\sum_{j=1}^{p} \hat{\theta}_{\hat{C}^{(g)}d,ij})$ for $r = 1, \ldots, R$. Finally, for each TF i and under each condition d, we compare the quantiles of the real data $\{X^{(g)}_{dr,i} : g = 1, \ldots, N; r = 1, \ldots, R\}$ to those of the pseudo-data $\{X^{*(g)}_{dr,i} : g = 1, \ldots, N; r = 1, \ldots, R\}$, which is summarized in Supplementary Material Tables S23–S52 [Luo and Wei (2018)]. From these tables, we can see that the marginal distribution of $X^{*(g)}_{dr,i}$ closely matches the distribution of $X^{(g)}_{dr,i}$, thus HDPGM fits the real data well.

6. Discussion. In this paper, we develop a rigorous statistical model, the HDPGM, to provide a legitimate measure of TF co-activation patterns across different cell types. On the one hand, the HDPGM simultaneously integrates information across biological conditions and the genome to improve signal detection from experiments with very few replicates. On the other hand, the HDPGM automatically teases out heterogeneity within each cell type, thus avoiding spurious associations, which can be very severe if a single network is assumed for all of the genomic locations. As a result, the HDPGM helps to achieve a valid and robust recovery of TF networks.

In our current model, the information is borrowed across the genome and biological conditions via the DP prior as discussed in Section 2.3. However, no further structure similarity is imposed for $\Theta_{k1}, \ldots, \Theta_{kd}, \ldots, \Theta_{kD}$ within a cluster. In principle, we can incorporate another layer of Markov random field prior to promote structure similarities across conditions as done in literature ignoring heterogeneity among samples [Lin et al. (2017), Mitra, Müller and Ji (2016), Peterson, Stingo and Vannucci (2015)]. We choose to focus on the current model here for the clarity of presentation. In addition, HDPGM can also consider the spatial correlations among genomic locations by employing the hidden Markov Dirichlet Process [Xing, Sohn et al. (2007)]. However, it will be much more computationally demanding. Fortunately, although HDPGM can lose some power in inferring the network structures when applied to datasets with spatial correlations, the power loss is not substantial, and HDPGM still provides good clusterings as discussed in Section 4.2.

Although thousands of ChIP-seq data have already been stored in public data repositories, such as ENCODE, we are still hungry for more ChIP-seq data to comprehensively study TF networks. The major obstacle is that the number of TFs whose ChIP-seq experiments are available for a large number of common biological conditions is still very small. Nevertheless, given the exponentially decreasing sequencing costs and statistical developments in imputation methods for genomic data [Ebert and Bock (2015)], TF networks with a larger number of nodes and conditions can be expected in the near future. As more data accumulate, our scalable HDPGM model will continue to update our understanding of TF co-activation, gene regulation, and ultimately phenotype diversity and diseases.

In addition to the TF association problem, many real-world networks are heterogeneous and dynamic while adopting the count data format. For example, the click numbers of news websites can be used to construct news networks that represent associations between different news categories. News networks are heterogeneous in relation to the readers' geographical locations, and they also change over time. We envision that our model will stimulate further statistical methodological research on heterogeneous dynamic networks for count data under the same framework. We also provide the C code implementing HDPGM on GitHub https://github.com/XiangyuLuo/HDPGM. More generally, we hope this study raises concerns about sample heterogeneity before blind application of all types of graphical models to real-world problems.

Acknowledgments. We thank the Editor, Associate Editor, and two reviewers for their invaluable and constructive comments which have greatly improved the quality of the paper. We also thank Jing Chu, a former undergraduate student at the Chinese University of Hong Kong, for her contribution to part of the C code for HDPGM.

SUPPLEMENTARY MATERIAL

Supplementary Materials to "Nonparametric Bayesian learning of heterogeneous dynamic transcription factor networks" (DOI: 10.1214/17-AOAS1129SUPP; .zip). The zip file provides the supplementary details referenced in the main text, the C code that implements HDPGM, and the datasets used in the simulation study and the real application.

REFERENCES

ALDOUS, D. J. (1985). *Exchangeability and Related Topics*. Springer, New York.

BICKEL, P. J. and LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969

BICKEL, P. J., BOLEY, N., BROWN, J. B., HUANG, H. and ZHANG, N. R. (2010). Subsampling methods for genomic inference. *Ann. Appl. Stat.* **4** 1660–1697.

CARTER, S. L., BRECHBÜHLER, C. M., GRIFFIN, M. and BOND, A. T. (2004). Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics* **20** 2242–2250.

CHENG, Y. and LENKOSKI, A. (2012). Hierarchical Gaussian graphical models: Beyond reversible jump. *Electron. J. Stat.* **6** 2309–2331.

CHENG, C., ALEXANDER, R., MIN, R., LENG, J., YIP, K. Y., ROZOWSKY, J., YAN, K.-K., DONG, X., DJEBALI, S., RUAN, Y. et al. (2012). Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res.* **22** 1658–1667.

CHUN, H., ZHANG, X. and ZHAO, H. (2015). Gene regulation network inference with joint sparse Gaussian graphical models. *J. Comput. Graph. Statist.* **24** 954–974.

DANAHER, P., WANG, P. and WITTEN, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 373–397.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **39** 1–38.

EBERT, P. and BOCK, C. (2015). Improving reference epigenome catalogs by computational prediction. *Nat. Biotechnol.* **33** 354–355.

ENCODE PROJECT CONSORTIUM (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489** 57–74.

ERNST, J. and KELLIS, M. (2012). ChromHMM: Automating chromatin-state discovery and characterization. *Nat. Methods* **9** 215–216.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

GAO, C., ZHU, Y., SHEN, X. and PAN, W. (2016). Estimation of multiple networks in Gaussian mixture models. *Electron. J. Stat.* **10** 1133–1154.

GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

GERSTEIN, M. B., KUNDAJE, A., HARIHARAN, M., LANDT, S. G., YAN, K.-K., CHENG, C., MU, X. J., KHURANA, E., ROZOWSKY, J., ALEXANDER, R. et al. (2012). Architecture of the human regulatory network derived from ENCODE data. *Nature* **489** 91–100.

GRANDORI, C., COWLEY, S. M., JAMES, L. P. and EISENMAN, R. N. (2000). The Myc/Max/Mad network and the transcriptional control of cell behavior. *Annu. Rev. Cell Dev. Biol.* **16** 653–699.

GROPP, W., LUSK, E. and SKJELLUM, A. (1999). *Using MPI*: *Portable Parallel Programming with the Message-Passing Interface*, *Vol.* 1. MIT Press, Cambridge, MA.

GUO, J., LEVINA, E., MICHAILIDIS, G. and ZHU, J. (2011). Joint estimation of multiple graphical models. *Biometrika* **98** 1–15.

GUO, J., LEVINA, E., MICHAILIDIS, G. and ZHU, J. (2015). Estimating heterogeneous graphical models for discrete data with an application to roll call voting. *Ann. Appl. Stat.* **9** 821–848. MR3371337

HANLEY, J. A. and MCNEIL, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143** 29–36.

HOBERT, O. (2008). Gene regulation by transcription factors and microRNAs. *Science* **319** 1785–1786.

INOUYE, D. I., YANG, E., ALLEN, G. I. and RAVIKUMAR, P. (2017). A review of multivariate distributions for count data derived from the Poisson distribution. *Wiley Interdiscip. Rev.*: *Comput. Stat.* **9** e1398, 25. MR3648601

ISHWARAN, H. and JAMES, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* **96** 161–173.

ISHWARAN, H. and RAO, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* **33** 730–773. MR2163158

JOHNSON, D. S., MORTAZAVI, A., MYERS, R. M. and WOLD, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316** 1497–1502.

KARLIS, D. (2003). An EM algorithm for multivariate Poisson distribution and related models. *J. Appl. Stat.* **30** 63–77.

KARLIS, D. and MELIGKOTSIDOU, L. (2007). Finite mixtures of multivariate Poisson distributions with application. *J. Statist. Plann. Inference* **137** 1942–1960.

KAWAMURA, K. (1979). The structure of multivariate Poisson distribution. *Kodai Math. J.* **2** 337–345.

KITAMURA, Y., SHIMOHAMA, S., OTA, T., MATSUOKA, Y., NOMURA, Y. and TANIGUCHI, T. (1997). Alteration of transcription factors NF-κB and STAT1 in Alzheimer's disease brains. *Neurosci. Lett.* **237** 17–20.

KOCHERLAKOTA, S. and KOCHERLAKOTA, K. (1992). *Bivariate Discrete Distributions*. Wiley, New York.

LAN, K.-H., KANAI, F., SHIRATORI, Y., OHASHI, M., TANAKA, T., OKUDAIRA, T., YOSHIDA, Y., HAMADA, H. and OMATA, M. (1997). In vivo selective gene expression and therapy mediated by adenoviral vectors for human carcinoembryonic antigen-producing gastric carcinoma. *Cancer Res.* **57** 4279–4284.

LARA-MARQUEZ, M. L., O'DORISIO, M. S., O'DORISIO, T. M., SHAH, M. H. and KARACAY, B. (2001). Selective gene expression and activation-dependent regulation of vasoactive intestinal peptide receptor type 1 and type 2 in human T cells. *J. Immunol.* **166** 2522–2530.

LI, S.-H. and LI, X.-J. (2004). Huntingtin–protein interactions and the pathogenesis of Huntington's disease. *Trends Genet.* **20** 146–154.

LIN, Z., WANG, T., YANG, C. and ZHAO, H. (2017). On joint estimation of Gaussian graphical models for spatial and temporal data. *Biometrics* **73** 769–779.

LOCHAMY, J., ROGERS, E. M. and BOSS, J. M. (2007). CREB and phospho-CREB interact with RFX5 and CIITA to regulate MHC class II genes. *Mol. Immunol.* **44** 837–847.

LUO, X. and WEI, Y. (2018). Supplement to "Nonparametric Bayesian learning of heterogeneous dynamic transcription factor networks." DOI:10.1214/17-AOAS1129SUPP.

MACARTHUR, S., LI, X.-Y., LI, J., BROWN, J. B., CHU, H. C., ZENG, L., GRONDONA, B. P., HECHMER, A., SIMIRENKO, L., KERÄNEN, S. V. et al. (2009). Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biol.* **10** R80.

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. and TELLER, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21** 1087–1092.

MITCHELL, P. J. and TJIAN, R. (1989). Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* **245** 371–378.

MITRA, R., MÜLLER, P. and JI, Y. (2016). Bayesian graphical models for differential pathways. *Bayesian Anal.* **11** 99–124. MR3447093

MITRA, R., MÜLLER, P., LIANG, S., YUE, L. and JI, Y. (2013). A Bayesian graphical model for chip-seq data on histone modifications. *J. Amer. Statist. Assoc.* **108** 69–80.

NEWTON, M. A., NOUEIRY, A., SARKAR, D. and AHLQUIST, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5** 155–176.

OGATA, H., GOTO, S., SATO, K., FUJIBUCHI, W., BONO, H. and KANEHISA, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **27** 29–34.

PETERSON, C. B., STINGO, F. C. and VANNUCCI, M. (2015). Bayesian inference of multiple Gaussian graphical models. *J. Amer. Statist. Assoc.* **110** 159–174.

ROBINSON, M. D., MCCARTHY, D. J. and SMYTH, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26** 139–140.

RODRIGUEZ, A., LENKOSKI, A., DOBRA, A. et al. (2011). Sparse covariance estimation in heterogeneous samples. *Electron. J. Stat.* **5** 981–1014.

SCHERZER, C. R., GRASS, J. A., LIAO, Z., PEPIVANI, I., ZHENG, B., EKLUND, A. C., NEY, P. A., NG, J., MCGOLDRICK, M., MOLLENHAUER, B. et al. (2008). GATA transcription factors directly regulate the Parkinson's disease-linked gene α-synuclein. *Proc. Natl. Acad. Sci. USA* **105** 10907–10912.

SHI, Q., LE, X., ABBRUZZESE, J. L., WANG, B., MUJAIDA, N., MATSUSHIMA, K., HUANG, S., XIONG, Q. and XIE, K. (1999). Cooperation between transcription factor AP-1 and NF-κB in the induction of interleukin-8 in human pancreatic adenocarcinoma cells by hypoxia. *J. Interferon Cytokine Res.* **19** 1363–1371.

SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R., LANDER, E. S. et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102** 15545–15550.

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** 528–540.

WEI, Y. and WU, H. (2016). Measuring the spatial correlations of protein binding sites. *Bioinformatics* **32** 1766–1772.

XING, E. P., SOHN, K.-A. et al. (2007). Hidden Markov Dirichlet process: Modeling genetic inference in open ancestral space. *Bayesian Anal.* **2** 501–527.

XUE, W., KANG, J., BOWMAN, F. D., WAGER, T. D. and GUO, J. (2014). Identifying functional co-activation patterns in neuroimaging studies via Poisson graphical models. *Biometrics* **70** 812–822.

YANG, E., RAVIKUMAR, P. K., ALLEN, G. I. and LIU, Z. (2013). On Poisson graphical models. In *Advances in Neural Information Processing Systems* 1718–1726.

YANG, E., RAVIKUMAR, P., ALLEN, G. I. and LIU, Z. (2015). Graphical models via univariate exponential family distributions. *J. Mach. Learn. Res.* **16** 3813–3847.

YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35.

ZERVOS, A. S., GYURIS, J. and BRENT, R. (1993). Mxi1, a protein that specifically interacts with Max to bind Myc-Max recognition sites. *Cell* **72** 223–232.

ZHANG, B. and HORVATH, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4** Article17.

ZHOU, H., CHERUVANKY, A., HU, X., MATSUMOTO, T., HIRAMATSU, N., CHO, M. E., BERGER, A., LEELAHAVANICHKUL, A., DOI, K., CHAWLA, L. S. et al. (2008). Urinary exosomal transcription factors, a new class of biomarkers for renal disease. *Kidney Int.* **74** 613–621.

DEPARTMENT OF STATISTICS
THE CHINESE UNIVERSITY OF HONG KONG
G26 LADY SHAW BUILDING
SHATIN, NT
HONG KONG
E-MAIL: xyluo1991@gmail.com

DEPARTMENT OF STATISTICS
THE CHINESE UNIVERSITY OF HONG KONG
111 LADY SHAW BUILDING
SHATIN, NT
HONG KONG
E-MAIL: yweicuhk@gmail.com