BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 16 2014, pages 2263–2271
doi:10.1093/bioinformatics/btu303

Sequence analysis

Advance Access publication April 29, 2014

Detecting clustering and ordering binding patterns among transcription factors via point process models

Maria Cha and Qing Zhou*

Department of Statistics, University of California, Los Angeles, CA 90095, USA

Associate Editor: Inanc Birol

ABSTRACT

Motivation: Recent development in ChIP-Seq technology has generated binding data for many transcription factors (TFs) in various cell types and cellular conditions. This opens great opportunities for studying combinatorial binding patterns among a set of TFs active in a particular cellular condition, which is a key component for understanding the interaction between TFs in gene regulation.

Results: As a first step to the identification of combinatorial binding patterns, we develop statistical methods to detect clustering and ordering patterns among binding sites (BSs) of a pair of TFs. Testing procedures based on Ripley's K-function and its generalizations are developed to identify binding patterns from large collections of BSs in ChIP-Seq data. We have applied our methods to the ChIP-Seq data of 91 pairs of TFs in mouse embryonic stem cells. Our methods have detected clustering binding patterns between most TF pairs, which is consistent with the findings in the literature, and have identified significant ordering preferences, relative to the direction of target gene transcription, among the BSs of seven TFs. More interestingly, our results demonstrate that the identified clustering and ordering binding patterns between TFs are associated with the expression of the target genes. These findings provide new insights into co-regulation between TFs.

Availability and implementation: See 'www.stat.ucla.edu/~zhou/TFKFunctions/' for source code.

Contact: zhou@stat.ucla.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 27, 2013; revised on April 2, 2014; accepted on April 24, 2014

1 INTRODUCTION

Gene expression in high organisms is regulated by the combinatorial binding of multiple transcription factors (TFs) to the DNA sequence in upstream or other regulatory regions of the gene. A set of interactive TFs often shows complex binding patterns. Clusters of multiple TF binding sites (TFBSs) co-localized in a short DNA region form the so-called *cis*-regulatory modules (CRMs). More complicated patterns, such as ordering preference among binding sites (BSs) of different TFs in a regulatory region, may exist and encode hidden regulatory signals. Detecting combinatorial binding patterns and CRMs is thus an important step toward a comprehensive understanding of *cis*-regulation. Consequently, computational methods have been developed for this purpose based on different types of genomic data.

Early methods build probabilistic models for TFBSs and their occurrences in CRMs to identify clusters of TFBSs from DNA sequence data. For a recent review of these methods, see Hardison and Taylor (2012). When the TFs involved in a particular cellular condition are known, methods have been developed to detect clusters of motif matches appearing more frequently than expected given a background model for the DNA sequence (Berman *et al.*, 2002; Frith *et al.*, 2002; Halfon *et al.*, 2002; Markstein *et al.*, 2002; Rebeiz *et al.*, 2002). There are also *de novo* methods for the discovery of CRMs, such as Gupta and Liu (2005); Zhou and Wong (2004, 2007) for a few examples. Essentially, this group of methods finds clusters of TFBSs based on the enrichment of motif matches compared against a simple background model. Because motif signals are usually weak in high organisms, sophisticated binding patterns, such as the ordering and other location preferences among TFBSs, cannot be reliably detected from sequence data alone and thus are often not incorporated in these methods.

Recently, ChIP-Seq technology has been applied to generate genome-wide binding data for TFs in various cellular conditions. The accumulation of such binding data for multiple TFs in the same condition makes it possible to explore the relation among the binding of these TFs. For example, more than 10 TFs have available ChIP-Seq data in mouse embryonic stem cells (ESCs) (Chen *et al.*, 2008; Heng *et al.*, 2010; Marson *et al.*, 2008). One may want to find out whether the binding of two TFs is independent of each other and, if not, identify specific binding patterns suggested by the data. Such results may indicate how two or more TFs work together to regulate the expression of target genes and help detect novel regulatory signals. However, those motif-enrichment methods reviewed above are not directly applicable owing to the different data type and their simplified model assumptions. On the other hand, *ad hoc* methods have been used to define and detect clusters of BSs or ChIP-Seq peaks in recent studies. Chen *et al.* (2008) defined multiple TF binding loci by peaks within 100 bp. Lee and Zhou (2013) identified context-dependent co-regulators from co-bound regions constructed by ChIP-Seq peaks of two TFs with distance <50 bp. Ji *et al.* (2006) determined whether binding regions have a clustering tendency based on the empirical distribution of peak-to-peak distances. Orlov *et al.* (2009) found clusters of BSs by iteratively merging peaks within 100 bp and then constructed some test using Poisson approximation. Kazemian *et al.* (2013) identified inter-site spacing bias for some fixed range between 0 and 100 bp by a Fisher's exact test on the

*To whom correspondence should be addressed.

© The Author 2014. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

contingency table of site pair counts within or outside the range. One sees an arbitrary use of distance thresholds and simple empirical tests in finding clusters of BSs (or peaks) in existing work, without any consideration of ordering preference among TFBSs or the individual location distribution for BSs of a TF. Clearly, there is a pressing need for new statistical methods to infer combinatorial binding patterns from ChIP-Seq data.

In this work, we develop statistical methods to detect clustering and ordering patterns between BSs of two TFs from their ChIP-Seq data. We assume that under the null model the locations of BSs of a TF follow an inhomogeneous Poisson point process. Test statistics for clustering and ordering binding patterns are constructed based on Ripley's *K*-function (Dixon, 2002; Ripley, 1976), a widely used method for detecting point patterns in spatial statistics, and its generalizations to pairwise TF binding data. Our approach does not rely on a pre-specified distance threshold for the detection of clusters of BSs and can automatically find the distance level at which the clustering pattern, if exists, is most significant. We make use of the large number of BSs in ChIP-Seq data and demonstrate that the null distributions of the test statistics can be accurately approximated by large-sample theory. This eliminates the need for computationally intensive simulations and allows for efficient large-scale analysis.

2 METHODS

Our methods are designed for the discovery of binding patterns from ChIP-Seq data of two TFs, X and Y. In this work, we focus on the upstream $(-8K, 2K]$ region, relative to the transcription start site (TSS), of a gene, which is arguably the region that contains most *cis*-regulatory signals. We collect all upstream regions that contain at least one ChIP-Seq peak from each of the two TFs, as our goal is to infer the relation between the binding of them. The corresponding genes of these upstream regions are called the (common) target genes of the two TFs. To simplify notation, we regard the location at $-8K$ bp as the origin and choose the direction of gene transcription as the positive direction, so that an upstream region has coordinates $(0, L]$ $(L = 10K)$. This data collection procedure gives a meaningful definition of the ordering among BSs, relative to the direction of gene transcription. Suppose that x and y are the coordinates of two BSs $(x, y \in (0, L])$. We say x precedes y if $x < y$. For the sake of understanding, we first describe our models and methods assuming that the exact locations of BSs are given, and then discuss a computational approach that predicts the exact location of a BS from a ChIP-Seq peak.

The input dataset for a pair of TFs X and Y consists of N upstream regions, $R_1, \ldots, R_N$, and a set of BS locations on each region. For $r = 1, \ldots, N$, let $\mathbf{x}_r = (x_{r,i}, i = 1, \ldots, n_r)$ and $\mathbf{y}_r = (y_{r,j}, j = 1, \ldots, m_r)$ be the BS locations for X and Y in the region R_r, respectively. Because we assume the same model for each region, the index r may be suppressed when we describe the methods for a generic region.

2.1 Inhomogeneous Poisson processes

Under the null model, BS locations of a TF are modeled by an inhomogeneous Poisson point process independent of the binding of other TFs. A Poisson point process on the line is a model for the occurrence of points in a one-dimensional interval. Denote the interval of interest by $(0, L]$, which represents an upstream region in this work. A Poisson point process can be equivalently specified by a counting process for the number of points that have occurred in any subinterval $(a, b]$, denoted by $N(a, b]$. For an inhomogeneous Poisson process with intensity function $\lambda(x)$, $x \in (0, L]$, the count $N(a, b]$ follows a Poisson distribution with rate $\lambda_{ab} = \int_a^b \lambda(x)dx$, i.e.

$$P\{N(a, b] = z\} = \frac{(\lambda_{ab})^z}{z!} e^{-\lambda_{ab}}, \; z = 0, 1, 2, \ldots. \tag{1}$$

In addition, $N(a_1, b_1]$ and $N(a_2, b_2]$ are independent if the two subintervals do not overlap. Given the total number of points $N(0, L] = n$, the points (their locations) can be regarded as i.i.d. with density function

$$f(x) = \frac{\lambda(x)}{\lambda_{0L}} \tag{2}$$

for $x \in (0, L]$. This important property of an inhomogeneous Poisson process is most relevant to this study. So far we have been assuming that a region of interest is a continuous interval, but the upstream regions are composed of discrete base pairs. This technical difficulty can be resolved by assuming that the intensity function $\lambda(x)$ is piecewise constant: for $x_1, x_2 \in (k, k+1]$, where k is an integer, we always have $\lambda(x_1) = \lambda(x_2)$. Under this assumption,

$$\int_a^b \lambda(x)dx = \sum_{k=a+1}^{b} \lambda(k) \tag{3}$$

for two integers $a < b$.

The Poisson point process is a reasonable model for BS locations of a TF. First, when the considered region is large, such as $L = 10K$ in our case, observing a BS in a small interval $(x - \Delta, x]$ is a rare event with probability $\propto \lambda(x)\Delta$ and it is independent of the occurrence of BSs in other non-overlapping intervals. Second, the inhomogeneity allows us to capture the common observation that BSs tend to occur more frequently in regions closer to the TSS. Therefore, we expect $\lambda(x)$ to have peaks when x is close to the TSS.

We make a key assumption that the conditional density $f(x)$ in (2) for a particular TF is identical across all regions. Denote by $f_X(x)$ the density for TF X. By Equation (2), the intensity function of TF X for the r^{th} region

$$\lambda_X^{(r)}(x) = \lambda_{X,0L}^{(r)} f_X(x), \tag{4}$$

where $\lambda_{X,0L}^{(r)} = \int_0^L \lambda_X^{(r)}(x)dx$. Note that the expected number of BSs of TF X in the region R_r is $\lambda_{X,0L}^{(r)}$. Thus, our null model accounts for the fact that each upstream region may have a different number of observed BSs, but assumes that the location distribution of the BSs is identical with a common density $f_X(x)$, which can be reliably estimated with enough regions. The intensity function of TF Y is defined in the same way as (4). From Equation (1) with $a = 0$ and $b = L$, one sees that $\lambda_{X,0L}^{(r)}$ can simply be estimated by n_r and $\lambda_{Y,0L}^{(r)}$ by m_r for $r = 1, \ldots, N$. To estimate f_X and f_Y from BS locations $(\mathbf{x}_r, \mathbf{y}_r)$, we further assume that the densities are piecewise constant where the length of a piece is h bp. Denote the pieces by B_i for $i = 1, \ldots, L/h$. Then $f_X(x)$ for $x \in B_i$ is estimated by

$$\hat{f}_X(x) = \frac{\text{total number of BSs for } X \text{ in } B_i \text{ over all regions}}{h \cdot n_\bullet}, \tag{5}$$

where $n_\bullet = \sum_{r=1}^{N} n_r$ is the total number of BSs for X. Consequently, the intensity function for the r^{th} region is estimated by

$$\hat{\lambda}_X^{(r)}(x) = n_r \hat{f}_X(x) \tag{6}$$

for TF X, and the same estimation is applied for TF Y.

2.2 Clustering detection

Ripley's *K*-function (Dixon, 2002; Ripley, 1976) is a broadly used analysis tool for summarizing and detecting point patterns in spatial data. In this

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

section, we develop a method based on a bivariate *K*-function for detecting clustering binding patterns between two TFs.

For two point processes, the bivariate *K*-function is the expected number of pairs of points, one from each process, with distance $\leq t$, normalized by the product of the intensities of the two processes (Baddeley *et al.*, 2000). Consider a generic upstream region with BS locations $\mathbf{x}=(x_1,\ldots,x_n)$ for TF X and $\mathbf{y}=(y_1,\ldots,y_m)$ for TF Y. An estimated *K*-function for this region is

$$\hat{K}_{XY}(t)=\frac{1}{L}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{w(x_i,y_j)^{-1}I(d_{x_i,y_j}\leq t)}{\lambda_X(x_i)\lambda_Y(y_j)}, \tag{7}$$

where $I(\cdot)$ is the indicator function, $w(x_i,y_j)$ is a weight function for edge correction and $d_{x_i,y_j}=|x_i-y_j|$ is the distance between x_i and y_j. Ignoring the edge correction at this moment, if there are many pairs of points within distance t relative to the intensities, $\hat{K}_{XY}(t)$ will be greater than its expected value under the assumption that the binding of the two TFs is independent of each other. Reversely, if the BSs of the two TFs are mutually repulsive, one would expect fewer pairs having distance $\leq t$ and thus a small $\hat{K}_{XY}(t)$. By varying the distance t over a wide range, we will obtain a curve of the bivariate *K*-function to summarize the clustering/repulsive pattern for all distance levels. This is the overall idea behind our method.

There are a few technical issues about the estimated *K*-function (7). First, edge effects may arise. There may be points lying outside the study area, $(0, L]$, even though they are within the distance t from a point located in the study area. Thus, to avoid bias in estimating the *K*-function, an appropriate edge correction should be applied to data points close to the boundaries of the region. Several weight functions have been suggested for edge correction (Yamada and Rogerson, 2003), and Ripley's method (Ripley, 1976) is used in this work. In a plane, $w(x_i,y_j)$ is defined as the fraction of the circumference of a circle centered at x_i with radius d_{x_i,y_j} that lies inside the study area. In the one-dimensional case where x_i and y_j are located on the line, the weight function reduces to

$$w(x_i,y_j)=\begin{cases}1 & \text{if } [x_i-d_{x_i,y_j},\,x_i+d_{x_i,y_j}]\subseteq(0,L]\\ 1/2 & \text{otherwise.}\end{cases} \tag{8}$$

Second, since the intensities λ_X and λ_Y are unknown, they are estimated by $n\hat{f}_X$ and $m\hat{f}_Y$ according to Equation (6).

To detect a potential clustering pattern, we calculate the *K*-function for each region, $\hat{K}^{(r)}_{XY}(t)$, given the BS locations $\mathbf{x}_r$ and $\mathbf{y}_r$. To evaluate its significance, the expectation and variance of $\hat{K}^{(r)}_{XY}(t)$ are calculated under the null hypothesis $\mathcal{H}_c$ that BSs of the two TFs are independently distributed according to inhomogeneous Poisson point processes. For any distance $t\in(0,L]$, conditional on the numbers of BSs, n_r and m_r,

$$\mathbb{E}\left[\hat{K}^{(r)}_{XY}(t)\mid n_r,m_r,\mathcal{H}_c\right]=2t, \tag{9}$$

$$\begin{aligned}&\mathrm{Var}\left[\hat{K}^{(r)}_{XY}(t)\mid n_r,m_r,\mathcal{H}_c\right]\\&=\frac{V(t)+4(m_r-1)C_Xt^2+4(n_r-1)C_Yt^2}{L^2n_rm_r}\triangleq[\sigma_r(t)]^2,\end{aligned} \tag{10}$$

where $V(t)$, C_X and C_Y can be calculated by numerical summation. See Supplementary Document for the detailed derivation. Note that $V(t)$, C_X and C_Y do not depend on r and thus are identical across different upstream regions, which makes the variance calculation efficient even for a large number of regions.

If the two TFs tend to bind in a cluster of distance t, observed value of $\hat{K}^{(r)}_{XY}(t)$ will be substantially greater than its expected value $2t$ for many regions. Conversely, if it is substantially smaller than $2t$, we may conclude that the two TFs have a significant repulsive binding pattern. Because we are interested in such overall binding patterns, it is natural to construct a test statistic that combines the *K*-functions from all regions. To this end, a normalized *Z*-score of the *K*-functions is defined by

$$Z_c(t)=\frac{1}{\sqrt{N}}\sum_{r=1}^{N}\frac{\hat{K}^{(r)}_{XY}(t)-2t}{\sigma_r(t)}, \tag{11}$$

where each term in the summation has zero mean and unit variance under $\mathcal{H}_c$. Thus, Z_c is the sum of independent random variables with identical means and variances (although their distributions may differ). By Lindberg's central limit theorem, which only requires independence among random variables,

$$Z_c(t)\mid\mathcal{H}_c\sim\mathcal{N}(0,1),\ \text{as } N\to\infty. \tag{12}$$

This leads to a simple null distribution for our test statistic when N is large. The accuracy of this asymptotic approximation will be evaluated for the data in this work by simulation. So far, we have been focusing on the procedure for a given t. In practice, we consider a range of values of $t\in(0,5000]$. This range is expected to be wide enough to capture all interesting clustering distances.

The testing procedure for detecting clustering patterns is summarized as follows.

(1) Estimate by Equations (5) and (6) the intensity functions of the two TFs, $\lambda^{(r)}_X(x)$ and $\lambda^{(r)}_Y(y)$, for all r.

(2) For each value of t:

(a) Calculate the $V(t)$, C_X and C_Y in (10).

(b) Calculate $\hat{K}^{(r)}_{XY}(t)$ and $\sigma_r(t)$ for $r=1,\ldots,N$.

(c) Calculate $Z_c(t)$ and its *P*-value by normal approximation.

Our method also reports the most significant clustering distance t^* that maximizes $Z_c(t)$, i.e. $t^*=\mathrm{argmax}_t Z_c(t)$.

2.3 Ordering detection

For a pair of TFs with significant clustering binding for some distance level, we further test whether their binding also has an ordering preference relative to the direction of target gene transcription. For a generic upstream region, let $\mathbf{X}=(X_1,\ldots,X_n)$ and $\mathbf{Y}=(Y_1,\ldots,Y_m)$ be two random vectors that represent the respective locations of the BSs of the two TFs, X and Y, showing the most significant clustering pattern at distance t^*. Our ordering detection method is based on modeling the conditional distribution of Y_i given $\mathbf{X}$. We assume that Y_i is generated independently given *one and only one* X_j from a conditional distribution, $P(Y_i=y|X_j=x)$, for $y=x-t^*,\ldots,x+t^*$, and call X_j the parent of Y_i. As y ranges between $x-t^*$ and $x+t^*$, we are only considering those Y_i such that $|Y_i-X_j|\leq t^*$ for some j. This makes the ordering analysis meaningful and well-defined. Hereafter, n and m, or n_r and m_r for the r^{th} region, refer to the numbers of BSs satisfying this distance constraint. Because there may be more than one X_j whose distance to a particular Y_i is $\leq t^*$, we introduce hidden variables Z_{ij} to indicate which X_j is the parent of Y_i, i.e.

$$Z_{ij}=\begin{cases}1 & \text{if } X_j \text{ is the parent of } Y_i,\\ 0 & \text{otherwise.}\end{cases} \tag{13}$$

We parametrize this model by $\theta=(\theta_{-t^*},\ldots,\theta_{t^*})$ with

$$\theta_d=P(Y_i=x+d\mid Z_{ij}=1,X_j=x)\ \text{for } d=-t^*,\ldots,t^*, \tag{14}$$

subject to the constraints that $\sum_d\theta_d=1$ and $\theta_d\geq 0$. With this parameterization, the null hypothesis that the binding of the two TFs has no ordering preference is expressed as

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

$$\mathcal{H}_o : \sum_{d<0} \theta_d = \sum_{d>0} \theta_d. \tag{15}$$

That is, Y_i has an equal chance to be located to the left or to the right of its parent. Moreover, we assume that *in priori* every X_j has an equal likelihood to be the parent of Y_i.

Because the model is essentially a mixture model with hidden variables (Z_{ij}), an EM algorithm is developed to estimate θ under $\mathcal{H}_o$ (Supplementary Document). Denote by $\hat{\theta}$ the estimate of θ output from the EM algorithm. Before we use this estimate for detecting ordering patterns, we apply some simple method to smooth $\hat{\theta}_d$ over d, as many spikes often exist in $\hat{\theta}$ because of having no or few observations for some distances d.

Let $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_m)$ be the observed binding locations of the two TFs on a generic upstream region R. We propose a new statistic $\hat{K}_{Y|X}(t^*)$ for ordering detection,

$$\hat{K}_{Y|X}(t^*) = \frac{1}{Lm} \sum_{i=1}^{m} \frac{c(y_i, \mathbf{x})}{f_{Y|X}(y_i|\mathbf{x})}, \tag{16}$$

where

$$c(y_i, \mathbf{x}) = \sum_{j=1}^{n} [I(x_j < y_i \le x_j + t^*) - I(x_j - t^* \le y_i < x_j)] \tag{17}$$

and $f_{Y|X}(y|\mathbf{x})$ denotes the conditional probability of $Y_i = y$ given $\mathbf{x}$ for any i (as they are i.i.d. given $\mathbf{x}$). Under our model assumptions,

$$f_{Y|X}(y|\mathbf{x}) = \frac{1}{n} \sum_{j \in J_{\mathbf{x}}(y)} \theta_{(y - x_j)}, \tag{18}$$

where $J_{\mathbf{x}}(y) = \{j : |x_j - y| \le t^*\}$. Under the null hypothesis $\mathcal{H}_o$, the expectation and variance of $\hat{K}^{(r)}_{Y|X}(t^*)$ for a region R_r, $r = 1, \ldots, N$, are

$$\mathbb{E}\left[\hat{K}^{(r)}_{Y|X}(t^*)|\mathbf{x}_r, m_r, \mathcal{H}_o\right] = 0, \tag{19}$$

$$\begin{aligned} &\mathrm{Var}\left[\hat{K}^{(r)}_{Y|X}(t^*)|\mathbf{x}_r, m_r, \mathcal{H}_o\right] \\ &= \frac{1}{L^2 m_r} \int_{A(\mathbf{x}_r)} \frac{[c(y, \mathbf{x}_r)]^2}{f_{Y|X}(y|\mathbf{x}_r)} dy \triangleq [\tau_r(t^*)]^2, \end{aligned} \tag{20}$$

where $A(\mathbf{x}_r) = \bigcup_{j=1}^{n_r} U_{r,j}$ and $U_{r,j} = [x_{r,j} - t^*, x_{r,j} + t^*] \cap (0, L]$ for $j = 1, \ldots, n_r$. See Supplementary Document for the derivation of these results.

We did not include edge correction in the definition (16) of $\hat{K}_{Y|X}(t^*)$. This is because the distance t^* for most clustering pairs is much smaller than L, the length of the upstream regions. For the significant ordering pairs we detected in this work, the percentage of BSs located within t^* bp to either boundary of the upstream regions is around or <1%. Thus, the edge effect is largely ignorable for the ordering detection method.

Finally, by the same argument as in the clustering analysis, a normalized Z-score is used as the test statistic for ordering detection,

$$Z_o(t^*) = \frac{1}{\sqrt{N}} \sum_{r=1}^{N} \frac{\hat{K}^{(r)}_{Y|X}(t^*)}{\tau_r(t^*)}, \tag{21}$$

which approximately follows the standard normal distribution $\mathcal{N}(0, 1)$ under $\mathcal{H}_o$ when N is large. From the definition of $c(y_i, \mathbf{x})$ in Equation (17), $Z_o(t^*) > 0$ implies that X binding precedes Y.

Now we summarize the testing procedure for ordering detection.

(1) Estimate $\theta = (\theta_{-t^*}, \ldots, \theta_{t^*})$ from all regions using the EM algorithm and smooth $\hat{\theta}$.

(2) For $r = 1, \ldots, N$, calculate $\hat{K}^{(r)}_{Y|X}(t^*)$ and $\tau_r(t^*)$, where $\hat{\theta}$ is plugged into Equation (18) to estimate the mixture density $f_{Y|X}(y|\mathbf{x}_r)$.

(3) Calculate $Z_o(t^*)$ and its P-value by normal approximation.

Obviously, one may switch the roles of X and Y in this procedure, modeling the conditional distribution of $\mathbf{X}$ given $\mathbf{Y}$. In practice, we calculate two P-values using the above procedure, one from $\hat{K}_{Y|X}$ and the other from $\hat{K}_{X|Y}$, and then take the geometric average of the two P-values as the combined evidence for ordering patterns.

2.4 Predicting BS locations

A ChIP-Seq peak for a TF is recorded as an interval in which a TFBS is supposed to locate. We need to pinpoint the exact BS locations from ChIP-Seq data before applying our methods for clustering and ordering detection. To do this, we first extract the DNA sequence for a ChIP-Seq peak using the software *CisGenome* (Ji *et al.*, 2008). Then, we scan the sequence, as well as its reverse compliment, with a sliding window of length w, where w is the length of the position-specific weight matrix (PWM) of the TF from the database TRANSFAC (Matys *et al.*, 2003). We score a w-mer in the DNA sequence by a likelihood ratio,

$$LR_i = \frac{P(s_i|\Theta)}{P(s_i|\theta_0)}, \tag{22}$$

where s_i is the nucleotide sequence of the i^{th} w-mer, Θ is the PWM for this TF and θ_0 is a background Markov model. The w-mer that has the maximum score among all w-mers scanned from the sequence is chosen as the exact BS, whose location (start position) is used in our analysis.

3 RESULTS

In this work, we study binding patterns in ChIP-Seq data generated from mouse ESCs. ESCs are pluripotent stem cells derived from the inner cell mass of the blastocyst, an early-stage embryo (Thomson *et al.*, 1998). They are known to have two distinctive properties, pluripotency and the ability to replicate indefinitely (Pan and Thomson, 2007), which make ESCs especially important in clinical research. It is expected that many complex diseases might be treated by transplanting cells generated from ESCs (Odorico *et al.*, 2001).

There has been a substantial amount of recent work on the gene regulatory network in mouse ESCs, which involves the combinatorial regulation by a number of TFs (Chen *et al.*, 2008; Ivanova *et al.*, 2006; Kim *et al.*, 2008; Zhou *et al.*, 2007). For example, Oct4 is known as a master regulator of pluripotency, and a valid amount of Oct4 is needed for dedifferentiation and for sustaining ESC self-renewal (Pan *et al.*, 2002). Four TFs, Oct4, Sox2, cMyc and Klf4, can reprogram somatic cells back to ESC-like cells having the characteristics of self-renewal and pluripotency (Takahashi and Yamanaka, 2006). Genome-wide ChIP-Seq data have provided the binding regions of >10 TFs that play key regulatory roles in mouse ESCs (Chen *et al.*, 2008; Heng *et al.*, 2010; Marson *et al.*, 2008). Computational analyses have been performed to identify the sequence motifs of these core TFs (Bailey, 2011; Mason *et al.*, 2010; Thomas-Chollier *et al.*, 2012) and to detect other co-regulators that may regulate genes by working with these TFs (Chen and Zhou, 2011; He *et al.*, 2009; Lee and Zhou, 2013). However, except for some *ad hoc* analysis on the co-occurrence between binding peaks, it appears that no principled approaches have been applied to detect combinatorial binding patterns from this rich collection of data. In this analysis, we apply our methods to 14 TFs with available ChIP-Seq data from three recent publications (Chen *et al.*,

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

2008; Heng *et al.*, 2010; Marson *et al.*, 2008) and systematically identify clustering and ordering binding patterns between all 91 TF pairs. The 14 TFs are Oct4, Sox2, Smad1, Stat3, Nanog, Esrrb, Tcfcp2l1, Tcf3, Nr5a2, E2f1, nMyc, cMyc, Zfx and Klf4. We also suggest potential regulatory roles of the identified combinatorial binding patterns by examining the expression profiles and functional annotations of their target genes.

We first verified via a simulation study that the normal approximations to the null distributions of the test statistics, Z_c (11) and Z_o (21), are accurate for the data analyzed in this work. One key feature that distinguishes our methods from other *ad hoc* methods is the use of inhomogeneous Poisson processes in modeling BS locations, which takes into account the location preference of TFBSs in upstream regions. In the simulation study, we also compared our approach against a simple method that ignores the location preference in detecting clustering. The simple method produced substantially more false discoveries. This shows that modeling the marginal binding distribution of a TF is a critical component for the detection of significant binding patterns. See Supplementary Document for a detailed description of the simulation study.

3.1 General clustering patterns

In this section, we analyze the potential clustering patterns of the 91 TF pairs. Because Oct4 and Sox2 are the most well-known and important master regulators in ESCs, we illustrate the details in each step of our analysis with this pair of TFs.

The analysis starts with estimating the density function f_X for each TF. To construct a robust estimate, we chose the bin size h so that every bin has at least one BS in at least one region. Then, we estimated the density function f_X by (5) and the intensity function $\lambda_X^{(r)}(x)$ by (6) for each region. As expected, both TFs tend to bind more frequently near the TSS of a gene (Supplementary Fig. S1). Such marginal location preferences are captured by the estimated densities. As demonstrated by our simulation study, ignoring this fact would lead to false positives in clustering detection.

With the estimated intensity functions for a pair of TFs, the bivariate K-function (7) for each upstream region and the Z-score (11) were calculated for a sequence of distances. See Figure 1(A) for the K-function of Oct4 and Sox2, averaged over upstream regions bound by the two TFs, which lies far above the expected value $2t$. This suggests that the two TFs have a significant clustering pattern for all distances, which is confirmed by the large positive values of $Z_c(t)$ in Figure 1B. The maximal Z_c value, 66.9, occurs at the distance $t^* = 50$, and thus the clustering between Oct4 and Sox2 binding is found most significant when examining the BSs from the two TFs located within 50 bp. The monotone decreasing trend of $Z_c(t)$ as t increases (Fig. 1B) shows that the clustering pattern becomes less significant as the distance threshold increases. This result is clearly consistent with the known tight clustering between the BSs of Oct4 and Sox2. In fact, in some regions, the two TFs bind adjacent sites with only a few (or even no) gaps (Chen *et al.*, 2008; Mason *et al.*, 2010).

To summarize the results of all 91 TF pairs, we consider the average of $Z_c(t)$ over $t \in [50, 1000]$, denoted by $\overline{Z}_{1000}$, for P-value and false discovery rate (FDR) calculations. First, using $\overline{Z}_{1000}$ is more conserved than using the maximal Z-score $Z_c(t^*)$ because the latter will cause a selection bias and artificially increase the significance level. Second, for most applications in defining clusters of BSs or CRMs, the maximal distance threshold is often a few hundred base pairs. Given $\overline{Z}_{1000}$ we calculated the P-value from $\mathcal{N}(0, 1)$ and the corresponding FDR for all TF pairs. It turns out that 86 out of the 91 pairs were found to have significant clustering patterns with an FDR $< 2\%$ (Supplementary Table S1). This is in line with the finding that most of these TFs show extensive co-binding in mouse ESCs (Chen *et al.*, 2008; Marson *et al.*, 2008). The statistics $\overline{Z}_{1000}$ of all the TF pairs are shown as a heatmap in Figure 2. For almost all significant pairs, the most significant clustering distance t^* is within 1000 bp (Supplementary Fig. S2), confirming short-range co-occurrence among BSs in regulatory regions. We further analyzed 11 TF pairs with $t^* = 50$ bp, as these TFs are likely to cluster at closer distances. After calculating their $Z_c(t)$ for $t < 50$, we found seven pairs having most significant clustering within 6 bp and the other four pairs close to 50 bp (Supplementary Table S2). BSs with a gap less than a few base pairs are often considered a composite site bound by a complex, while those separated by a few hundred base pairs have been seen in CRMs (Zhou and Wong, 2004).

We compared our approach with a simple overlapping analysis for the co-occurrence of BSs of two TFs in the upstream regions. We calculated the χ^2 statistics and found that all TF pairs but one (Tcf3 and cMyc) are significant. The χ^2 statistics and our $\overline{Z}_{1000}$ are positively correlated, but there is no exact correspondence between the two (Supplementary Fig. S3). Our method gives more information than the simple overlapping analysis does. It not only suggests that the TFBSs co-occur in upstream regions but also provides the significance levels for different distances. Furthermore, our method is able to detect a clustering pattern where two TFs bind a certain base pairs

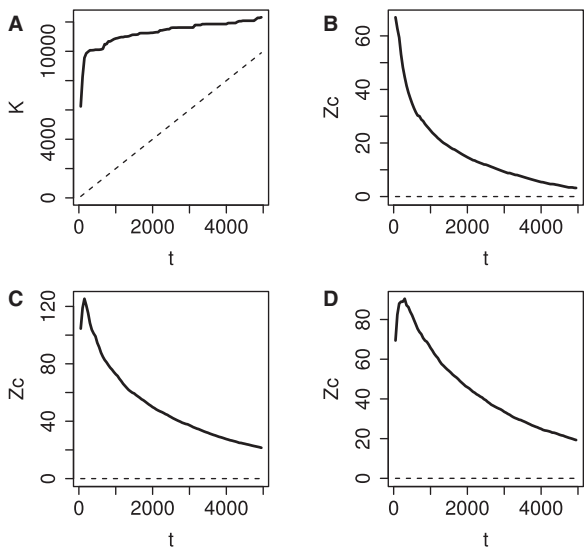

Fig. 1. Example clustering binding patterns: (**A**) bivariate K-function of Oct4 and Sox2 (solid line), averaged over all regions, and the expected value $2t$ (dashed line); (**B**) $Z_c(t)$ of Oct4 and Sox2 ($t^* = 50$); (**C**) $Z_c(t)$ of nMyc and E2f1 ($t^* = 150$); and (**D**) $Z_c(t)$ of Zfx and E2f1 ($t^* = 300$)

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

apart, reflected by a peak in the curve of $Z_c(t)$, such as the examples shown in Figure 1C and D.

Our clustering statistic $\overline{Z}_{1000}$ can provide detailed information about the co-binding pattern among a set of TFs. To demonstrate this, we performed single-linkage hierarchical clustering of the 14 TFs with $\overline{Z}_{1000}$ as the pairwise similarity measure. The left margin of the heatmap in Figure 2 shows the resulting cluster dendrogram, in which TFs that merge at a lower level have a more significant clustering pattern. One sees that a tight cluster can be formed by cMyc, nMyc, E2f1 and Zfx with high values of $\overline{Z}_{1000}$ between them. These four TFs are the members of the so-called cMyc group defined in previous studies (Chen *et al.*, 2008; Ouyang *et al.*, 2009). Further moving up the dendrogram, the cMyc group merges with Klf4 and other TFs from the Oct4 group, including Oct4, Sox2, Nanog, Esrrb, Tcfcp2l1, Smad1 and Stat3. As expected, one sees another tight cluster of Oct4, Sox2 and Nanog from the heatmap. In addition to partitioning the TFs into two separate groups as suggested by Chen *et al.* (2008), our result reveals that some Oct4 group TFs extensively co-bind with the cMyc group as well. The existence of a large number of genes extensively bound by both groups has been reported in previous studies (Chen *et al.*, 2008; Chen and Zhou, 2011). An integrated analysis of ChIP-Seq, motif and gene expression data suggested that the co-binding among the Oct4 group TFs occurs in a Myc-dependent way (Lee and Zhou, 2013). In both ESCs and induced pluripotent stem cells, genes associated with cMyc binding, by itself or in combination with Oct4, Sox2 and Klf4, are significantly enriched for regulators of metabolic processes, while genes co-bound by Oct4, Sox2 and Klf4 in absence of cMyc are mainly implicated in regulation of development (Sridharan *et al.*, 2009). Combined with these published data, the tree structure in Figure 2 provides new biological insights with interesting functional implications into the co-regulation among these core TFs in mouse ESCs. We also noticed that Tcf3, a component in the Wnt-signaling pathway, showed

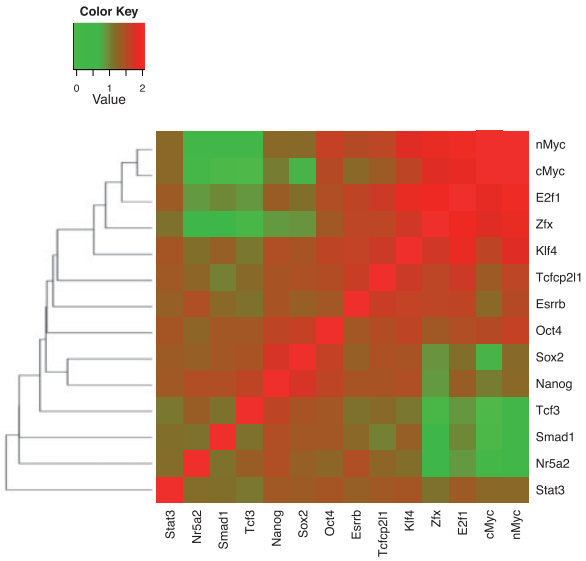

Fig. 2. Heatmap with the hierarchical clustering tree for the statistics $\log_{10}(\overline{Z}_{1000})$ between the 14 TFs

significant clustering patterns with the core ESC regulators, Oct4, Sox2 and Nanog, consistent with the observation in Marson *et al.* (2008). It was verified in a recent study (Zhang *et al.*, 2013) by electrophoretic mobility shift assay that Tcf3 competes with Sox2 for the Sox motif in Oct-Sox composite sites, which may counter pluripotency. This shows another example of complicated interactions among multiple TFs to implement diverse transcriptional programs. These examples support the view that combinatorial binding patterns may have important implications in the expression and functions of target genes. Thus, we performed a systematic combined analysis of clustering binding patterns and gene expression data (see next subsection).

3.2 BS clustering and differential gene expression

Zhou *et al.* (2007) have generated gene expression profiles in mouse ESCs by sorting cells according to the expression level of the gene Oct4. The expression data consist of eight samples with high Oct4 expression and eight samples with low Oct4 expression. By two-sample comparison between the two groups of samples, they defined 1325 Oct4-sorted+ genes, which showed a strong positive correlation with Oct4 expression, and 1440 Oct4-sorted– genes with the opposite expression pattern. It is interesting to examine whether the clustering binding pattern between a pair of TFs is different between the upstream regions of these two gene sets. To give a concrete example, among the 375 common target genes of Oct4 and Sox2, 56 of them are in the Oct4-sorted+ set and 11 in the Oct4-sorted– set. We wish to test if the clustering binding pattern of these two TFs shows any difference between the upstream regions of the 56 Oct4-sorted+ genes and those of the 11 Oct4-sorted– genes. In general, for a pair of TFs X and Y, denote by S_1 and S_2, respectively, the sets of Oct4-sorted+ and Oct4-sorted– target genes of the two TFs. For each gene in either set, we calculate a Z-score for a distance level t,

$$Z_c^{(r)}(t) = \frac{\hat{K}_{XY}^{(r)}(t) - 2t}{\sigma_r(t)} \quad \text{for } r \in S_1 \cup S_2, \tag{23}$$

which is one term in the summation of (11). Note that every $Z_c^{(r)}(t)$ has zero mean and unit variance under the null hypothesis $\mathcal{H}_c$. Now regarding $\{Z_c^{(r)}(t) : r \in S_1\}$ and $\{Z_c^{(r)}(t) : r \in S_2\}$ as two groups of standardized observations, we use the two-sample t-statistic, denoted by $T(t)$, to test if the two group means are identical. Again, an overall statistic, $\overline{T}_{1000}$, is calculated as the mean of $T(t)$ over $t = 50, \ldots, 1000$ for each pair.

Among the 86 TF pairs showing a clustering pattern, 15 of them also have a significant $\overline{T}_{1000}$ at an FDR of $<5\%$ (Supplementary Table S3). The significant $\overline{T}_{1000}$ statistics are all positive, indicating that the mean of the $Z_c^{(r)}$ statistics of the Oct4-sorted+ target genes is greater than that of the Oct4-sorted– target genes. In other words, the clustering binding pattern between a TF pair in the table is more significant in upstream of the Oct4-sorted+ genes than in that of the Oct4-sorted– genes. One possible explanation for this difference is that these TFs work in concert to activate Oct4-sorted+ genes, which are upregulated in ESCs, but may regulate Oct4-sorted– genes and promote early differentiation with other co-regulators not included in our analysis. This result demonstrates a clear association between the co-binding pattern of two TFs and the

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

expression of their target genes, suggesting that how TFs co-localize to the DNA sequence may serve as some subtle signal for downstream gene regulation.

3.3 Ordering among TF binding

We apply the ordering detection method to the 86 pairs of TFs with a significant clustering pattern. To better understand the detailed procedure, we describe each step of our analysis for the pair of E2f1 (X) and Zfx (Y).

For this pair, the most significant clustering distance $t^* = 300$, and thus, the model parameter $\theta = (\theta_{-300}, \ldots, \theta_{300})$ as defined in (14). We counted the number of pairwise BSs, x_j and y_i, with distance $d = 0, \ldots, t^*$ and added a little pseudo-count to each value of d to avoid zero counts. These counts were normalized into frequencies c_d, $d = 0, \ldots, t^*$, which were used as the initial estimate $\theta^{(0)}$ with $\theta_0^{(0)} = c_0$ and $\theta_{-d}^{(0)} = \theta_d^{(0)} = c_d/2$ for $d \geq 1$. As a result, the initial estimate of θ was symmetric. Starting with $\theta^{(0)}$, the EM algorithm developed in the Supplementary Document was run until convergence. We then smoothed the EM estimate $\hat{\theta}$ by a cubic smoothing spline (using the R function 'smooth.spline'). The plots for the initial value $\theta^{(0)}$ and the EM estimate $\hat{\theta}$ (after smoothing) for this pair are shown in Supplementary Figure S4, with a demonstration of the mixture density $f_{Y|X}(y|\mathbf{x})$ (18). Note that although the initial estimate is symmetric by construction, the EM estimate is not symmetric but satisfies the constraint (15) specified by the null hypothesis $\mathcal{H}_o$. With this way of initialization, the EM algorithm often converges after a few iterations. Given $\hat{\theta}$, it is easy to follow the ordering testing procedure to calculate $Z_o(t^*)$. For E2f1 and Zfx, we obtained $Z_o(300) = 12.0$ with a P-value of 0. Clearly, there is a significant ordering binding pattern between the two TFs within a distance of $t^* = 300$ bp. Moreover, the positive Z_o value indicates that the BSs of Zfx (Y) are much more likely to locate to the right of those of E2f1 (X), along the direction of gene transcription. We also repeated the same procedure but switched the labeling of the two TFs to see if the result is consistent. The resulting ordering statistic $Z_o' = -12.7$, which suggests the same ordering between the binding of the two TFs with an almost identical significance level.

At an FDR of $< 1\%$, we detected seven pairs of TFs showing ordering preferences between their BSs (Table 1). For all these pairs, switching the labels of X and Y gave essentially the same conclusion on the ordering pattern with comparable absolute values between the two statistics, Z_o and Z_o'. The geometric mean of the P-values of Z_o and Z_o' was calculated as the overall measure of the statistical significance for a pair. Note that in the table all pairs are labeled so that X precedes Y (i.e. $Z_o > 0$) for easy understanding. The ordering between two TFs may be a local feature that only occurs for BSs within the clustering distance t^* or a global feature that reflects that the BSs of one TF tend to locate in upstream of those of the other one. Both patterns can be captured by our testing procedure. We performed a two-sample comparison on the BS locations for each of the seven pairs and found that three of them, E2f1/cMyc ($P = 7.4 \times 10^{-15}$), E2f1/Zfx ($P = 7.2 \times 10^{-11}$) and Tcfcp2l1/Zfx ($P = 3.8 \times 10^{-5}$), showed significant global difference in their binding locations consistent with the orderings reported in Table 1. The other four pairs showed no significant global difference and thus the detected ordering patterns here are specific to tightly clustered BSs.

The pairwise ordering patterns can be combined to produce an ordering relation graph among multiple TFs. Regarding cMyc and nMyc together as the Myc TF, we encapsulate all the detected pairwise ordering patterns into such a graph in Figure 3A. In this graph, an edge from X to Y means that X precedes Y along the direction of gene transcription. If X precedes Y and Y precedes Z, we conclude that X also precedes Z and link the three TFs into a chain, $X - Y - Z$. One sees three chains ending at Zfx, meaning that the BSs of Zfx tend to be the closest one to the TSS. In addition, there is a full ordering among the cMyc group TFs (E2f1, cMyc/nMyc and Zfx) and Klf4, which is often regarded as a TF in between the Oct4 and the cMyc groups (Ouyang *et al.*, 2009) (also see Fig. 2). A example region with the exact full ordering is provided in Figure 3B. This seems to suggest a special sequential ordering in the co-binding of the cMyc group TFs. We found 866 genes whose upstream regions have BSs from all the three cMyc group TFs within 300 bp, the maximum t^* value between pairs in this group (Table 1), and 233 of them have at least one 300-bp region in which the BSs follow the exact ordering (E2f1–Myc–Zfx). We call the 233 genes the ordering group (Supplementary Table S4) and the other 633 genes the non-ordering group. Enriched gene ontology terms with P-value $< 10^{-9}$ in the ordering group include ribosome biogenesis, translation, macromolecule metabolic process and RNA

Table 1. TF pairs with a significant ordering pattern

X / Y	t^*	$Z_o(t^*)$	$Z_o'(t^*)$	P-value	FDR
E2f1/Zfx	300	11.991	–12.738	0	0
Klf4/Zfx	500	9.582	–7.623	0	0
nMyc/Zfx	250	4.869	–4.743	1.54×10^{-6}	4.40×10^{-5}
E2f1/cMyc	200	4.030	–3.632	1.25×10^{-4}	2.69×10^{-3}
Klf4/E2f1	200	3.726	–3.810	1.64×10^{-4}	2.83×10^{-3}
Oct4/Zfx	350	4.326	–2.847	2.59×10^{-4}	3.71×10^{-3}
Tcfcp2l1/Zfx	300	3.299	–3.573	5.85×10^{-4}	7.19×10^{-3}

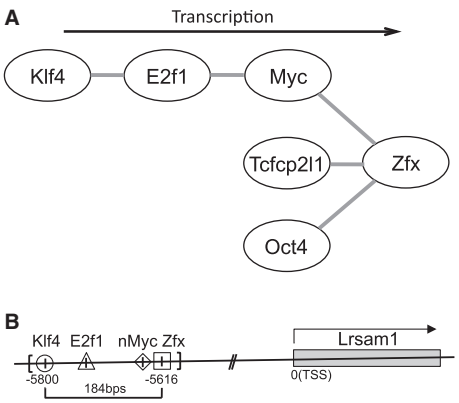

Fig. 3. Summary and an example of detected ordering patterns: (**A**) ordering relation graph among six TFs; (**B**) an example target gene (Lrsam1) and its upstream region with the exact ordering pattern among Klf4 and the cMyc group TFs as shown in (A)

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

processing, suggesting involvement in some most basic and fundamental biological processes. There are 28 Oct4-sorted+ but only 5 Oct4-sorted– genes in the ordering group, showing a noticeable depletion of genes downregulated in ESCs. We then examined the expression profiles in the Oct4-sorted series (Zhou *et al.*, 2007) of the ordering group (Supplementary Fig. S5) and those of the non-ordering group (Supplementary Fig. S6). The expression profile of each gene was normalized to have zero mean and unit standard deviation. The ordering group has a higher normalized expression level in the Oct4-high samples than the non-ordering group ($P = 8.0 \times 10^{-5}$), while the non-ordering group shows a higher expression level in the Oct4-low samples than the ordering group ($P = 1.3 \times 10^{-5}$). These observations indicate potential association between the ordering among BSs and the expression pattern of the target gene. Further experimental investigations hold the key to understanding the regulatory role of the exact ordering among the BSs and its relation to the functions of the target genes.

4 DISCUSSION

We have developed new statistical methods to detect combinatorial binding patterns from ChIP-Seq data. More specifically, we have proposed test statistics for clustering and ordering analysis for a pair of TFs. To construct flexible and threshold-free statistics, bivariate *K*-functions for Poisson point processes are used and further developed. With a proper normalization, the null distributions can be well-approximated by the standard normal distribution, which avoids the use of simulation-based *P*-value approximation. This makes our approach suitable for large-scale data. The test statistics can be used not only to make a binary decision, but also to measure how strong a specific binding pattern is between two TFs. Our procedures also offer simple ways to extend the pairwise analysis to multiple TFs, as demonstrated by the hierarchical tree among the binding patterns (Fig. 2) and the ordering relation graph (Fig. 3A). Novel and interesting biological insights can be provided by such new analysis. To the best of our knowledge, our method is the first systematical development for detecting ordering patterns among TFBSs. This work reports strong evidence (with small *P*-values and FDRs) for the existence of specific binding ordering among a set of TFs.

In this study, we predict the exact BSs from ChIP-Seq peaks by scoring *w*-mers, assuming that each ChIP-seq peak has exactly one BS. The maximum-scored *w*-mer is regarded as the BS for the region. One future direction is to take into account the uncertainty in the exact location of a BS. For example, we may normalize the likelihood ratio scores (22) to define a probability distribution for the BS in a ChIP-Seq peak and incorporate this distribution into our detection procedures. We may also consider the possibility that some peaks are false positives and may not contain a BS. In this case, exact *P*-value calculation (Touzet and Varré, 2007) will be needed to decide which *w*-mers are more likely the BSs. A direct generalization to multivariate *K*-functions seems infeasible when the number of TFs becomes large. To extend our methods to the study of multiple TF binding patterns, we may analyze the change in pairwise patterns with respect to the binding of a third TF or some score constructed by the binding of a few other TFs.

Funding: This work was supported by NSF [grants DMS-1055286 and DMS-1308376 to Q.Z.].

Conflict of Interest: none declared.

REFERENCES

Baddeley,A.J. *et al.* (2000) Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Stat. Neerlandica*, **54**, 329–350.

Bailey,T.L. (2011) DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, **27**, 1653–1659.

Berman,B.P. *et al.* (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *PNAS*, **99**, 757–762.

Chen,X. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.

Chen,G. and Zhou,Q. (2011) Searching ChIP-seq genomic islands for combinatorial regulatory codes in mouse embryonic stem cells. *BMC Genomics*, **12**, 515.

Dixon,P.M. (2002) Ripley's K function. *Encyclopedia Environmetrics*, **3**, 1796–1803.

Frith,M.C. *et al.* (2002) Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res.*, **30**, 3214–3224.

Gupta,M. and Liu,J.S. (2005) *De novo* cis-regulatory module elicitation for eukaryotic genomes. *PNAS*, **102**, 7079–7084.

Halfon,M.S. *et al.* (2002) Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. *Genome Res.*, **12**, 1019–1028.

Hardison,R.C. and Taylor,J. (2012) Genomic approaches towards finding cis-regulatory modules in animals. *Nat. Rev. Genet.*, **13**, 469–483.

He,X. *et al.* (2009) A biophysical model for analysis of transcription factor interaction and binding site arrangement from genome–wide binding data. *PLoS One*, **4**, e8155.

Heng,J.D. *et al.* (2010) The nuclear receptor Nr5a2 can replace Oct4 in the reprogramming of Murine somatic cells to Pluripotent cells. *Cell Stem Cell*, **6**, 167–174.

Ivanova,N. *et al.* (2006) Dissecting self-renewal in stem cells with RNA interference. *Nature*, **442**, 533–538.

Ji,H. *et al.* (2006) A comparative analysis of genome-wide chromatin immunoprecipitation data for mammalian transcription factors. *Nucleic Acids Res.*, **34**, e146.

Ji,H. *et al.* (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.*, **26**, 1293–1300.

Kazemian,M. *et al.* (2013) Widespread evidence of cooperative DNA binding Widespread evidence of cooperative DNA binding by transcription factors in *Drosophila* development. *Nucleic Acids Res.*, **41**, 8237–8252.

Kim,J. *et al.* (2008) An extended transcriptional network for pluripotency of embryonic stem cells. *Cell*, **132**, 1049–1061.

Lee,Y. and Zhou,Q. (2013) Co-regulation in embryonic stem cells via context-dependent binding of transcription factors. *Bioinformatics*, **29**, 2162–2168.

Markstein,M. *et al.* (2002) Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. *PNAS*, **99**, 763–768.

Marson,A. *et al.* (2008) Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell*, **134**, 521–533.

Mason,M.J. *et al.* (2010) Identification of context-dependent motifs by contrasting ChIP binding data. *Bioinformatics*, **26**, 2826–2832.

Matys,V. *et al.* (2003) TRANSFAC (R): transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.

Odorico,J.S. *et al.* (2001) Multilineage differentiation from human embryonic stem cell lines. *Stem Cells*, **19**, 193–204.

Orlov,Y.L. *et al.* (2009) Genome-wide statistical analysis of multiple transcription factor binding sites obtained by ChIP-seq technologies. In: *Proceedings of the 1st ACM Workshop on Breaking Frontiers of Computational Biology (CompBio'09)*. New York, 2009.

Ouyang,Z. *et al.* (2009) ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *PNAS*, **106**, 21521–21526.

Pan,G.J. *et al.* (2002) Stem cell pluripotency and transcription factor Oct4. *Cell Res.*, **12**, 321–329.

Pan,G. and Thomson,J.A. (2007) Nanog and transcriptional networks in embryonic stem cell pluripotency. *Cell Res.*, **17**, 42–49.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Rebeiz,M. *et al.* (2002) SCORE: A computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. *PNAS*, **99**, 9888–9893.

Ripley,B.D. (1976) Second-order analysis of stationary point processes. *J. Appl. Probability*, **13**, 255–266.

Sridharan,R. *et al.* (2009) Role of the Murine reprogramming factors in the induction of Pluripotency. *Cell*, **136**, 364–377.

Takahashi,K. and Yamanaka,S. (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, **126**, 663–676.

Thomas-Chollier,M. *et al.* (2012) RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res.*, **40**, 1–9.

Thomson,J.A. *et al.* (1998) Embryonic stem cell lines derived from human blastocysts. *Science*, **282**, 1145–1147.

Touzet,H. and Varré,JS. (2007) Efficient and accurate p-value computation for position weight matrices. *Algorithms Mol. Biol.*, **2**, 15.

Yamada,I. and Rogerson,P.A. (2003) An empirical comparison of edge effect correction methods applied to K-function analysis. *Geogr. Anal.*, **35**, 97–109.

Zhang,X. *et al.* (2013) Gene regulatory networks mediating canonical Wnt signal directed control of pluripotency and differentiation in embryo stem cells. *Stem Cells*, **13**, 2667–2679.

Zhou,Q. and Wong,W.H. (2004) CisModule: *de novo* discovery of cis-regulatory modules by hierarchical mixture modeling. *PNAS*, **101**, 12114–12119.

Zhou,Q. and Wong,W.H. (2007) Coupling hidden Markov models for the discovery of cis-regulatory modules in multiple spicies. *Ann. Appl. Stat.*, **1**, 36–65.

Zhou,Q. *et al.* (2007) A gene regulatory network in mouse embryonic stem cells. *PNAS*, **104**, 16438–16443.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016