View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Collection Of Biostatistics Research Archive

Harvard University

Harvard University Biostatistics Working Paper Series

Year 2011 *Paper* 133

Multiple Testing of Local Maxima for Detection of Peaks in ChIP-Seq Data

Armin Schwartzman* Andrew Jaffe†

Yulia Gavrilov‡ Clifford A. Meyer**

*Harvard School of Public Health and Dana Farber Cancer Institute, armin@jimmy.harvard.edu

†Johns Hopkins Bloomberg School of Public Health

‡Dana-Farber Cancer Institute

**Johns Hoplins Bloomberg School of Public Health

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

http://biostats.bepress.com/harvardbiostat/paper133

Copyright ©2011 by the authors.

Multiple Testing of Local Maxima for Detection of Peaks in ChIP-Seq Data

Armin Schwartzman[1,2], Andrew Jaffe[3], Yulia Gavrilov[1,2], Clifford A. Meyer[1,2]

[1,2]Department of Biostatistics, Harvard School of Public Health, and
[1,2]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute
[3]Department of Epidemiology and Department of Biostatistics,
Johns Hopkins Bloomberg School of Public Health

August 3, 2011

Abstract

A topological multiple testing approach to peak detection is proposed for the problem of detecting transcription factor binding sites in ChIP-Seq data. After kernel smoothing of the tag counts over the genome, the presence of a peak is tested at each observed local maximum, followed by multiple testing correction at the desired false discovery rate level. Valid p-values for candidate peaks are computed via Monte Carlo simulations of smoothed Poisson sequences, whose background Poisson rates are obtained via linear regression from a Control sample and the local GC content. The proposed method resolves nearby binding sites that other methods do not.

1 Introduction

The problem of detecting signal peaks in the presence of background noise appears often in the analysis of high-throughput data. In ChIP-Seq data, the problem of finding transcription factor binding sites along the genome translates to a large-scale peak detection problem with a one-dimensional spatial structure, where the number, locations and heights of the peaks are unknown. Recently, Schwartzman et al. (2010) (hereafter SGA) introduced a topological multiple testing approach to peak detection where, after kernel smoothing, the presence of a signal is tested not at each spatial location but only at the local maxima of the smoothed observed sequence. In this paper, we show how that approach can be used to formalize the inference problem of finding binding sites in ChIP-Seq data. To achieve this, we also propose a new regression-based method for estimating the local background binding rate from a Control sample.

1.1 ChIP-Seq data

ChIP-Sequencing or ChIP-Seq is an experimental method that is used to map the locations of binding sites of transcription factors along the genome in vivo Barski and Zhao (2009); Park (2009). Transcription factors control the transcription of genetic information from DNA to mRNA in living cells, and abnormalities in this process are often associated with cancer. Given a particular transcription factor of interest, ChIP-Seq combines chromatin immunoprecipitation

Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

(ChIP) with massively parallel DNA sequencing, allowing enrichment of the DNA segments bound by the transcription factor and mapping of their locations along the genome. The result is a long list of sequenced forward and reverse tags, also called reads, each associated with a specific genomic address. After alignment of these tags, the data consists of a sequence of tag counts along the genome, with a tendency to a higher concentration of tags near the transcription factor binding sites. An example of a data fragment is shown in Rows 1 and 2 of Figure 1.

The goal of the analysis is to identify the true binding sites. This translates to finding genomic locations where the binding rate is higher than it would be if the transcription factor were not present. To this end, Johnson et al. (2007) suggested sequencing a Control input sample to provide an experimental assessment of the background tag distribution, helping reduce false positives. The cost currently associated with this technology often does not allow more than a single ChIP-Seq sample, also called IP sample, and a single Control sample. To illustrate the usefulness of the Control, Rows 1 and 2 of Figure 1 show a short fragment of the raw data after alignment in the Control and IP samples respectively for the same positions in the genome. The interesting peaks are marked by red circles in Row 3, corresponding to sites with high binding rate in the IP sample but lower rate in the Control. Other candidate peaks, marked in blue, do not have a significantly higher binding rate in the IP sample than in the Control.

As an additional condition, it is not enough that a site has higher binding rate relative to the Control, but it must have a high binding rate in absolute terms. In other words, sites with a very weak binding rate are not interesting, even if the corresponding binding rate in the Control is even weaker.

1.2 Testing of local maxima

The search for binding sites may be set up as large-scale multiple testing problem where, at each genomic location, a test is performed for whether the binding rate is higher than the background. Testing at each genomic location is statistically inefficient because it requires a multiple testing correction for a very large number of tests over the entire length of the genome. In ChIP-Seq, the binding rate at a true binding site has a unimodal peak shape that spreads into neighboring locations, caused by the variability in the length of the sequenced segments. Thus, as argued by SGA, it is enough to test for high binding rates only at locations that resemble peaks, that is, local maxima of the smoothed data. In this sense, the local maxima serve as topological representatives of the candidate binding sites.

Peak detection on the aligned data is carried out using the Smooth and Test Local Maxima (STEM) algorithm of SGA. It consists of:

1. kernel smoothing;
2. finding the local maxima as candidate peaks;
3. computing p-values for the heights of the observed local maxima; and
4. applying a multiple testing procedure to the obtained p-values.

For Step 1, following the 'matched filter principle' recommended by SGA, we use a symmetric unimodal kernel that roughly matches the shape of the peaks to be detected. This shape corresponds to the spatial spread of tag locations around a true binding site and is assumed to be the same for all binding sites, up to an amplitude scaling factor, as dictated by the

COBRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper133

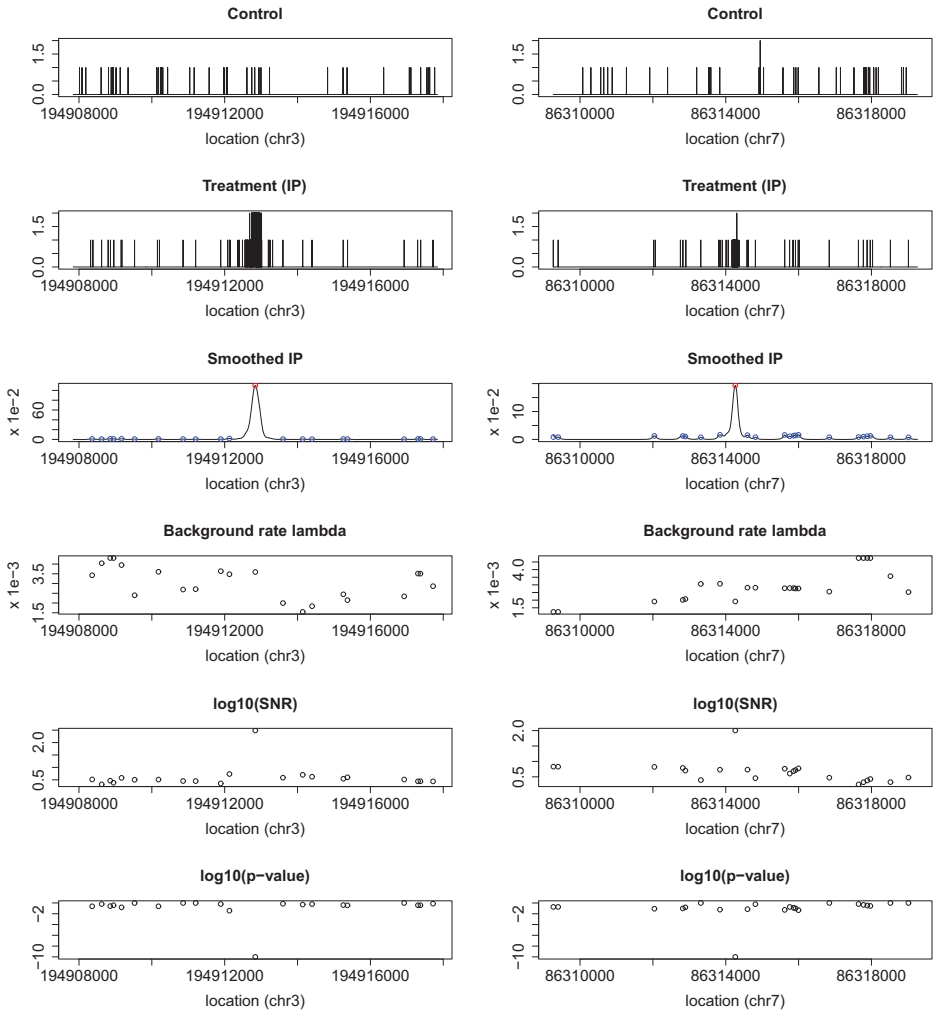


Figure 1: A fragment of the Fox A1 aligned data featuring a few representative peaks found by our method. Row 1: Control sample. Row 2: IP sample, same fragment as the Control. Row 3: Smoothed IP sample; significant peaks are indicated in red, non-significant ones in blue. Row 4: Estimates of the background Poisson rate $\lambda_0(t)$ at local maxima of the smoothed IP sample. Row 5: Signal-to-noise ratio (SNR), equal to peak height divided by background rate (log 10 scale). Row 6: P-values (log 10 scale)). Notice the difference in vertical scales between the left and right panels.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

physics and chemistry experimental protocol. This shape, up to an amplitude scaling factor, is estimated from the data during the alignment process. In Step 2, local maxima are defined as smoothed counts that are higher than their neighbors after correcting for ties. In Step 3, p-values test the hypothesis that the local binding rate is less or equal to the local background rate or a minimally interesting binding rate. The required distribution of the heights of local maxima is computed via Monte Carlo simulations. Finally, Step 4 is carried out using the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995), although in general, other multiple testing algorithms may be used instead.

The STEM algorithm is promising for ChIP-Seq data because it was shown in SGA to provide asymptotic error control and power consistency under similar modeling assumptions. Like in ChIP-Seq data, SGA assumed that the signal peaks are unimodal with finite support and that the search occurs over a long observed sequence. Under those conditions, and assuming the noise to be additive Gaussian stationary ergodic, SGA proved that the BH procedure controls the false discovery rate (FDR) of detected peaks. Here, FDR is defined as the expected ratio of falsely detected peaks among detected peaks, where a detected peak is considered true (false) if it occurs inside (outside) the support of any true peak. The control is asymptotic as both the search space and the signal strength get large. SGA also showed that the statistical power of the algorithm tends to one under the same asymptotic conditions. In ChIP-Seq data, the definitions of true and false detected peaks apply within the spatial extent of the true peak shape, which is estimated here during the alignment process.

1.3 Estimation of the background rate and Monte Carlo calculation of p-values

ChIP-Seq data differs from the modeling assumptions of SGA in that ChIP-Seq data consists of a long sequence of Poisson positive integer counts (Mikkelsen et al., 2007), which cannot be modeled with Gaussian signal-plus-noise model. Moreover, the process generating the background noise counts is not globally stationary (Johnson et al., 2007). To make inference possible, we assume the Poisson rate to vary over the genome but not too fast so that it is approximately constant in the immediate vicinity of any candidate peak. The background Poisson rate at any given location is estimated as a linear function of the local Control counts at three different scales and a nonlinear function of the local GC content. The linear coefficients are estimated from the data by multiple regression, automatically solving the normalization problem of having different sequencing depths between the IP and Control samples.

Finally, as required by Step 3 of the STEM algorithm above, for an observed local maximum of the smoothed ChIP-Seq data at a given location, its p-value is computed via Monte Carlo simulation using the background Poisson parameter estimated for that location. Note that the STEM algorithm requires an estimate of the background, but does not depend on how that estimate was obtained. Here we propose a regression method, but that method could be easily changed without changing the basic operation of the STEM algorithm.

1.4 Other methods

A number of methods to analyze ChIP-Seq data have been proposed by others before. Examples include MACS (Zhang et al., 2008), cisGenome (Ji et al., 2008), QuEST (Valouev et al., 2008), and FindPeaks (Fejes et al., 2008). Some of these methods also view the problem of detecting binding sites as a peak detection problem. While these methods use statistical models and estimate error rates, they do not formally state the inference problem they attempt to solve from

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper133

a statistical point of view (an exception, that frames the problem from a Bayesian perspective, is BayesPeak (Spyrou et al., 2009); our approach is frequentist). Here we attempt to frame the ChIP-Seq analysis problem as a formal inference problem in multiple testing, relying on the error control properties proven in SGA, and using a regression method to estimate the background binding rate. As a reference, we compare the results of our analysis to those of MACS and cisGenome on two different datasets. A particular feature of our method is that it focuses on detecting binding sites rather than binding regions, and therefore has the ability to distinguish nearby binding sites that the other methods do not.

1.5 Datasets

We demonstrate our method on two different ChIP-Seq datasets. In the first, ChIP-Seq targetting the transcription factor FoxA1 was performed on the breast cancer cell line MCF-7 (Zhang et al., 2008). This dataset includes a ChIP-Seq sample (hereafter IP), in which the FoxA1 antibody was used, and a Control input sample, in which the procedure was repeated without the antibody. Sequencing covered the entire genome, producing about 3.9 million tags in the IP sample and about 5.2 million tags in the Control sample. The second dataset concerns the growth-associated binding protein (GABP) (Valouev et al., 2008). This larger dataset consists of an IP sample with about 7.8 million tags and a Control sample with about 17.4 million tags. The methods in this paper were developed on the FoxA1 dataset, and later applied to the GABP dataset as an independent testbed. In both datasets, the goal of the analysis is to detect genomic loci in the IP sample that have a significantly high numbers of tags both in absolute terms and relative to the Control sample.

The methods in this paper were implemented in `R`.

2 Peak detection for ChIP-Seq data

2.1 Alignment and estimation of the peak shape

Before statistical analysis, we follow the approach in MACS of first aligning the forward and reverse tags, after which that distinction is unnecessary and tags can be counted together. Alignment requires estimating the amount by which tags need to be shifted. This process, described in the detail in Appendix, also allows us to estimate the shape of the spatial spread of the shifted tag counts around a peak.

For illustration, Figure 2a shows the spatial distributions of the forward and reverse tags in the IP sample of the FoxA1 dataset before alignment, obtained from 1000 strong and easily detectable peaks in chromosome 1. These distributions are displaced with respect to one another. The optimal shift found in this case was 62 base pairs (bp), almost the same as the estimated shift of 63 found by MACS for the same data. Figure 2a shows the estimated peak shape after alignment of these distributions (black dashed).

As a further refinement, it can be observed in Figure 2a that the binding rate is approximately constant more than about 400 bp away from the center, and should not be included as part of the peak. As a correction, the support of the kernel was reduced by multiplying the black-dashed shape by a quartic biweight function of size $W = 801$, producing the estimate in solid black. This peak shape, normalized to unit sum, is used as a smoothing kernel in the STEM algorithm for peak detection. The quartic biweight function has the effect of providing the kernel with continuous derivatives at the edges, a desirable property to avoid spurious local maxima at that step of the algorithm.

Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

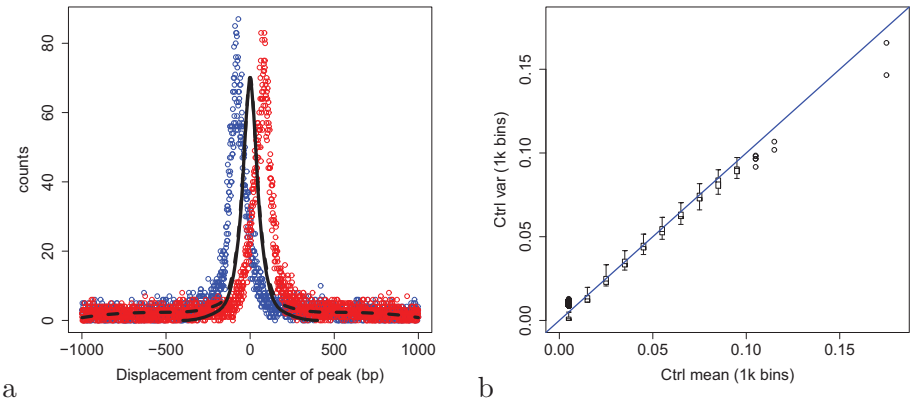


Figure 2: (a) Estimated distribution of tag counts in the forward strand (red) and in the reverse strand (blue) of the FoxA1 dataset (chromosome 1). Aligning the distributions and averaging the counts results in the joint count distribution and peak shape (black dashed). The peak shape is multiplied by a quartic biweight function (black solid). (b) Sample mean vs. sample variance of the aligned Control sequence in bins of size 1 Kb. The blue line has slope 1.

2.2 The Poisson model and the STEM algorithm

After alignment, the data consists of a table of genomic locations, each with an associated tag count. The rest of genomic locations are assumed to have a count of zero. Since the data are given as positive integer counts, it is reasonable to model them as Poisson variables (Mikkelsen et al., 2007). Specifically, we assume that the IP and Control counts $\mathrm{IP}(t)$ and $\mathrm{C}(t)$ at locations t are independent Poisson sequences

$$\mathrm{IP}(t) \sim Po\left[\lambda_{\mathrm{IP}}(t)\right], \qquad \mathrm{C}(t) \sim Po\left[\lambda_{\mathrm{C}}(t)\right], \qquad t \in \mathbb{Z}, \tag{1}$$

where $\lambda_{\mathrm{IP}}(t) \geq 0$ and $\lambda_{\mathrm{C}}(t) \geq 0$ denote the mean rates at location t, which may vary over t. The values of the processes $\mathrm{IP}(t)$ and $\mathrm{C}(t)$ are assumed independent over t given $\lambda_{\mathrm{IP}}(t)$ and $\lambda_{\mathrm{C}}(t)$.

As model validation, Figure 2b shows a graph of the sample mean vs. sample variance of the aligned Control sequence in the FoxA1 dataset, computed in bins of size 1 Kilobase (Kb). The two quantities are seen to be nearly proportional with a proportionality constant of 1, as expected from the Poisson model (1). The IP sample exhibits a similar pattern (not shown).

Regions of high binding frequency are represented by peaks in the mean Poisson rates. The goal is to find regions where $\lambda_{\mathrm{IP}}(t)$ is higher than the local background rate $\lambda_0(t)$, but also higher than a minimally interesting binding rate λ_{L}. The rate λ_{L} does not depend on t and establishes a lower bound on the height of peaks to be detected, ensuring that the detected peaks are strong not only relative to the local background but also in absolute terms. This avoids detecting uninteresting weak peaks where an even weaker local background would deem them statistically significant otherwise.

At every t, the above comparison translates to testing whether $\lambda_{\mathrm{IP}}(t) \leq \lambda_0(t)$ and $\lambda_{\mathrm{IP}}(t) \leq \lambda_{\mathrm{L}}$, that is, $\lambda_{\mathrm{IP}}(t) \leq \max\{\lambda_0(t), \lambda_{\mathrm{L}}\}$. To gain efficiency, rather than testing at every single location t, tests are only performed at local maxima of the smoothed IP sequence. This is carried out formally using the following adaptation of the STEM algorithm from SGA.

Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper133

Algorithm 1 (STEM algorithm)**.**

1. Let $w(t)$ be a unimodal kernel of length W. Apply kernel smoothing to the IP sequence to produce the smoothed sequence

$$\widetilde{\mathrm{IP}}(t) = w(t) * \mathrm{IP}(t) = \frac{1}{W} \sum_{s=-(W-1)/2}^{(W+1)/2} w(s)\mathrm{IP}(t-s). \tag{2}$$

2. Find all local maxima of $\widetilde{\mathrm{IP}}(t)$ as candidate peaks. Let $\tilde{T}$ denote the set of locations of those local maxima.

3. For each local maximum $t \in \tilde{T}$, compute a p-value $p(t)$ for testing the null hypothesis

$$\mathcal{H}_0(t) : \lambda_{\mathrm{IP}}(t) \leq \lambda_0^+(t) \qquad \text{vs.} \qquad \lambda_{\mathrm{IP}}(t) > \lambda_0^+(t) \tag{3}$$

   in a neighborhood of t, where $\lambda_0^+(t) = \max\{\lambda_0(t), \lambda_{\mathrm{L}}\}$.

4. Let $\tilde{m}$ be the number of local maxima. Apply a multiple testing procedure on the set of p-values and declare significant all peaks whose p-values are smaller than the threshold.

Details on each of the steps are given in the following sections.

2.3 Smoothing and local maxima

According to SGA, the best smoothing kernel for the purposes of peak detection is that which maximizes the signal-to-noise ratio (SNR) after convolving the peak shape, assumed to underly the signal peaks in the data, with the smoothing kernel. This is achieved by choosing the smoothing kernel to be equal to the peak shape itself (up to a scaling factor), a principle called "matched filter theorem" in signal processing (Pratt, 1991; Simon, 1995). Note that this is not the same as the optimal kernel in nonparametric regression.

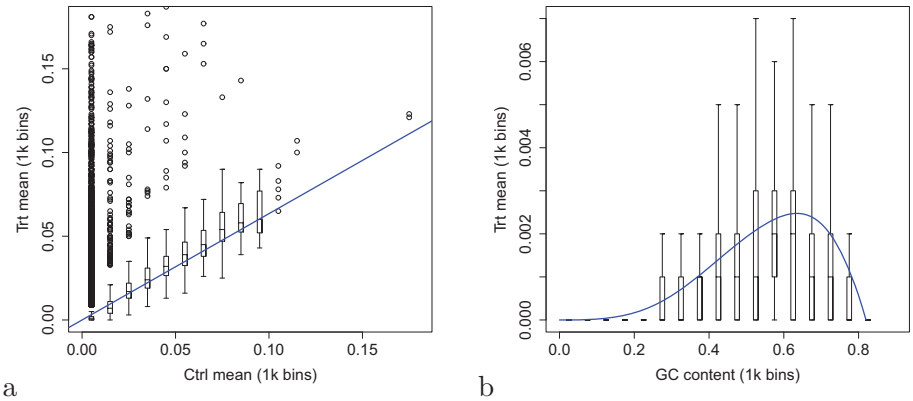
In ChIP-Seq data, binding rate peaks corresponding to different binding sites for the same transcription factor are assumed to have the same shape in terms of spatial spread, but may have different heights. The common peak shape is estimated in the alignment process (solid curve in Figure 2). It is unimodal, constrained to be symmetric, and has heavier tails than the Gaussian density. In Step 1 of the STEM algorithm (Algorithm 1), smoothing was carried out setting $w(t)$ equal to the solid curve in Figure 2, normalized to have unit sum, with $W = 801$.

Rows 2 and 3 in Figure 1 compare the raw and smoothed IP data. The smoothed data is high at locations where the density of tag counts is high. Notice that kernel smoothing produces positive counts locations where the unsmoothed IP data may have no counts.

In Step 2 of the STEM algorithm (Algorithm 1), local maxima of the smoothed sequence $\widetilde{\mathrm{IP}}(t)$ are defined as values $\widetilde{\mathrm{IP}}(t)$ that are greater than their immediate neighbors $\widetilde{\mathrm{IP}}(t-1)$ and $\widetilde{\mathrm{IP}}(t+1)$. If the maximum is tied between two neighboring values, then the peak location is assigned the mean position of the tied maxima, and the height their maximum smoothed value. A useful property of the kernel that avoids producing spurious local maxima is to have continuous derivatives. This was ensured by multiplication of the estimated peak shape by a quartic biweight function, as described in Section 2.1 above.

Restricting the analysis to local maxima reduces the amount of data to process further. In the aligned IP sample of the FoxA1 dataset, the number of local maxima found was about 2.7 million, down from about 3.9 million original mapped tags.

Hosted by The Berkeley Electronic Press



a b

Figure 3: Marginal distributions of the 1 Kb bin averages in the IP sample of the FoxA1 dataset as a function of: (a) the 1 Kb bin averages in the Control sample; (b) the 1 Kb bin GC content.

2.4 Estimation of the local background rate

Computation of p-values in Step 3 of the STEM algorithm (Algorithm 1) requires knowledge of the background Poisson rate $\lambda_0(t)$ under the null hypothesis. Estimation of $\lambda_0(t)$ is difficult because it varies with t in an unknown fashion (Johnson et al., 2007). Here we propose a simple method to estimate the background rate from the local Control data and the local GC content, as follows. The GC content is included as it is known to generally affect binding rates.

Since the Control sample is intended to represent the background process in the IP sample, it is reasonable to assume that the local background rate $\lambda_0(t)$ in the IP sample is a linear function of the corresponding local background rate $\lambda_C(t)$ in the Control sample. The proportionality accounts for the difference in sequencing depth between the two samples. In the FoxA1 data, the IP sample has about 3.9 million tags, while the Control sample has about 5.2 million counts.

Assuming that $\lambda_0(t)$ varies slowly with t, $\lambda_0(t)$ at a particular location t can be estimated as a linear function of the corresponding rate $\lambda_C(t)$ in the Control. The local Control rate $\lambda_C(t)$, in turn, may be estimated as the average tag count in the Control sample within a window of size 1 Kb centered at t. The size 1 Kb is about the smallest precision that allows comparison of peaks, usually of size a few hundred bp, against the background. Because counts may often be sparse, to add stability to the parameter estimates we consider the local rate to also be linearly related to the corresponding rate in the Control within a window of size 10 Kb and within a window of size 100 Kb, all centered at t.

To illustrate these relationships, Figure 3a shows a graph of the 1 Kb bin averages in the Control sample of the FoxA1 dataset against the 1 Kb bin averages in the IP sample. While there is a lot of variability, the main trend is seen to be linear, captured in the figure by a marginal linear fit. The outliers in the upper left corner correspond vaguely to the peaks sought. However, their relative small number introduces little bias in the regression. Similar trends are observed when plotting the 1 Kb bin averages in the IP sample as a function of the 10 Kb and 100 Kb bin averages in the Control sample (not shown).

Figure 3b shows the distribution of the 1 Kb bin averages in the IP sample of the FoxA1 dataset as a function of the GC content in the same 1 Kb bins. It is apparent that the local GC content affects the binding rates, but it does so in a nonlinear fashion. This relationship

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Table 1: Estimated regression coefficients by non-negative least squares in the FoxA1 dataset. Standard errors were computed for reference from ordinary least squares.

Predictor	1 Kb	10 Kb	100 Kb	GC spline				
Coefficient	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
Estimate	0.30378	0.28403	0.19787	0.00000	0.00000	0.00011	0.00036	0.00000
Std. error	0.00086	0.00204	0.00206	0.00004	0.00001	0.00001	0.00004	0.00009

may be captured nonparametrically using 5 B-spline basis functions.

Summarizing, $\lambda_0(t)$ is estimated from local windows of sizes 1 Kb, 10 Kb and 100 Kb centered at t via

$$\hat{\lambda}_0(t) = a_1\hat{\lambda}_{\mathrm{C},1k}(t) + a_2\hat{\lambda}_{\mathrm{C},10k}(t) + a_3\hat{\lambda}_{\mathrm{C},100k}(t) + \sum_{j=4}^{13} a_j b_j\left[\mathrm{GC}_{1k}(t)\right] \tag{4}$$

where $\hat{\lambda}_{\mathrm{C},1k}(t)$, $\hat{\lambda}_{\mathrm{C},10k}(t)$, and $\hat{\lambda}_{\mathrm{C},100k}(t)$ are the Control averages in windows of size 1 Kb, 10 Kb and 100 Kb centered at t, $\mathrm{GC}_{1k}(t)$ is the GC content in a window of size 1 Kb centered at t, $b_1(\cdot), \ldots, b_5(\cdot)$ are B-spline basis functions defined on the interval $(0, 0.8)$, and $a_1, \ldots, a_8$ are global parameters.

To estimate the global parameters $a_1, \ldots, a_8$, we set up a global linear regression as in (4), except that the predictors and the response are replaced by the 1 Kb, 10 Kb, and 100 Kb bin averages, as in Figure 3. The linear regression is solved with the additional constraint that all coefficients, and thus the fitted values, are non-negative.

Applying this regression in the FoxA1 dataset gave the estimates listed in Table 1. Interestingly, the three window sizes are given approximately equal weight, while the GC content is given very little weight. The coefficients also automatically account for sequencing depth. If the binding rate in the Control were constant, then the background estimate for the IP would be approximately equal to the Control rate multiplied by the sum of the three Control window coefficients, equal to 0.786. This factor is close to the overall ratio between the total number of reads in the IP sample and the total number of reads in the Control sample, equal to 0.747. However, the multi-window model makes the estimate adaptive to the local variability in the background rate.

As an example, Row 4 of Figure 1 shows the local estimates $\hat{\lambda}_0(t)$, roughly following the tag pattern observed in Row 1. Row 5 shows the SNR, defined as the ratio between the peak height $\widetilde{\mathrm{IP}}(t)$ and the estimated background rate $\hat{\lambda}_0(t)$.

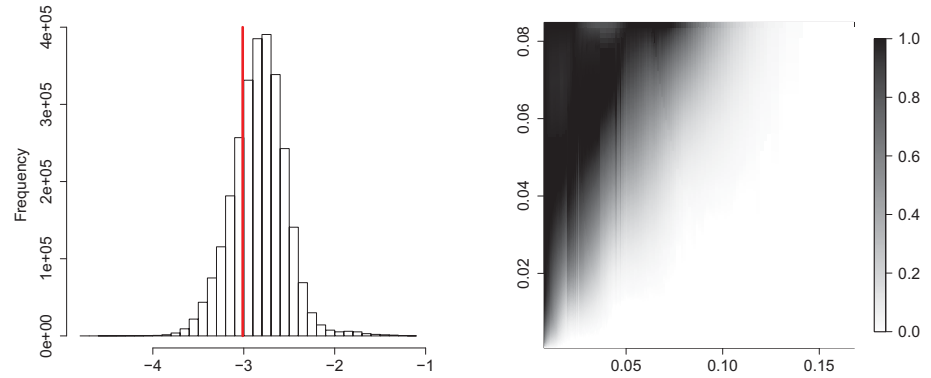
Figure 4a shows the distribution of the estimated values of $\hat{\lambda}(t)_0$ over the entire genome for the FoxA1 dataset. A reasonable value for the minimally interesting binding rate λ_{L} is the median of this distribution. We thus defined $\lambda_{\mathrm{L}} = \mathrm{median}_t\{\hat{\lambda}_0(t)\} = 0.00097$. This value corresponds to about 1 tag in a 1 Kb window, which is indeed low and uninteresting.

2.5 Computing p-values

In Step 3 of Algorithm 1, the p-value $p(t)$ of an observed local maximum of the smoothed sequence $\widetilde{\mathrm{IP}}(t)$ at a location t is defined as the probability to obtain the observed height of the local maximum or higher under the least favorable null hypothesis $\lambda_{\mathrm{IP}}(t) = \lambda_0^+(t)$ in (3) in a neighborhood of t. For the purposes of computing p-values, the null hypothesis need only be

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press



a b

Figure 4: (a) Histogram of estimated values of $\hat{\lambda}_0$ in the FoxA1 dataset. The median (marked in red) is 0.00097. (b) Right-tail distributions $\hat{F}(u;\lambda)$ of the height u of local maxima, obtained by Monte Carlo simulation, as a function of the background rate λ.

assumed in a local neighborhood of each candidate peak because $\widetilde{\mathrm{IP}}(t)$ depends only on the data within a local neighborhood, as dictated by the smoothing kernel $w(t)$. In this section we assume that $\lambda_0(t)$ and λ_{L} are known, having been estimated according to the methods described in Section 2.4 above.

In SGA, the background noise process was assumed stationary. In ChIP-Seq data, in contrast, the background rate $\lambda_0(t)$ is not constant. However, if the background process is locally stationary and $\lambda_0(t)$ varies slowly with t, then the background process in the neighborhood of a given location $t = \tilde{t}$ may be assumed to have similar statistical properties in that neighborhood as a stationary sequence with constant background rate $\lambda \equiv \lambda_0(\tilde{t})$. In particular, the height of a local maximum of the smoothed sequences at $\tilde{t}$ would have approximately the same distribution in both cases.

Specifically, suppose $X(t;\lambda)$ is a sequence of i.i.d. Poisson random variables with constant mean rate λ. Smoothing of $X(t;\lambda)$ with the kernel $w(t)$ as in Step 1 of Algorithm 1 produces the smoothed sequence

$$\tilde{X}(t;\lambda) = w(t) * X(t;\lambda) = \frac{1}{W} \sum_{s=-(W-1)/2}^{(W+1)/2} X(t-s;\lambda). \tag{5}$$

The height of a local maximum of the stationary sequence $\tilde{X}(t;\lambda)$ has the *right* cumulative distribution function (cdf)

$$F(u;\lambda) = \mathrm{P}\left[\tilde{X}(t;\lambda) \geq u \;\middle|\; t \text{ is a local maximum}, \lambda\right]. \tag{6}$$

Then, assuming that $\mathrm{IP}(t)$ is locally stationary, the required distribution of the height of a local maximum of $\widetilde{\mathrm{IP}}(t)$ at t may be approximated by the distribution $F(u;\lambda)$ (6) corresponding to the constant rate $\lambda \equiv \lambda_0(t)$. Finally, given the observed height $\widetilde{\mathrm{IP}}(t)$ at t, its p-value under the null hypothesis (3) is defined as

$$p(t) = F\left(\widetilde{\mathrm{IP}}(t); \lambda_0^+(t)\right). \tag{7}$$

Collection of Biostatistics Research Archive

http://biostats.bepress.com/harvardbiostat/paper133

The distribution (6) is difficult to compute analytically. Instead, we resort to Monte Carlo simulations, where for each given value of λ, a sequence $X(t;\lambda)$ of independent Poisson variables is generated as defined above smoothed using the kernel $w(t)$, and its local maxima found. The distribution (6) is then estimated empirically from the obtained heights of the local maxima of the smoothed simulated sequence $\tilde{X}(t;\lambda)$ (5).

To reduce computations, rather than performing a new simulation for each new background rate $\lambda_0(t)$, a table of cdfs (6) is prepared in advance for a set of values of λ that covers the range of possible values of $\lambda_0(t)$ to be found in the data. Then for any specific value of $\lambda_0(t)$, the table is interpolated to find the distribution (6) corresponding to that rate.

In the FoxA1 dataset, the smallest and largest values of $\hat{\lambda}_0(t)$ found were 1.67×10^{-5} and 7.49×10^{-2}, respectively, giving values of $\hat{\lambda}_0^+(t)$ in the range 0.00097 to 0.0749. Taking a safety margin of 25%, we performed the Monte Carlo simulation described above for 300 values of λ equally spaced on a logarithmic scale between 0.00073 and 0.0842. The length of the simulated Poisson sequences was set to be as long as needed to obtain at least 100 nonzero counts, but not smaller than 1×10^5. In order to reduce the variability from the simulation, the table of cdfs $\hat{F}(u,\lambda)$ was smoothed nonparametrically over λ for each fixed u via linear regression using 5 B-spline basis functions, with the additional constraint that all coefficients, and thus the fitted values, are non-negative. The safety margin mentioned above ensured that none of the values of λ actually neded were near the edges of the table for the purposes of spline smoothing.

Figure 4b shows the obtained function $\hat{F}(u;\lambda)$ (6), given as a table of size 300 values of λ by 200 values of u. In order to evaluate $\hat{F}(u;\lambda)$ in (7) for any particular values of $\widetilde{\mathrm{IP}}(t)$ and $\lambda_0(t)$, bilinear interpolation was used between the closest grid points.

As an example, Row 6 of Figure 1 shows the calculated p-values $p(t)$ in that data segment. Because of numerical precision in the Monte Carlo simulations, very low p-values could not be distinguished from zero. In the figures, these are drawn as if they were equal to 10^{-10}.

2.6 Multiple testing

Following SGA, we applied the BH procedure on the sequence of $\tilde{m} = 2683941$ p-values from the FoxA1 dataset, each corresponding to a local maximum of the smoothed sequence $\widetilde{\mathrm{IP}}(t)$. Of these local maxima, 17993 were declared significant at an FDR level of 0.01. Their associated addresses t are effectively point estimates of the locations of the binding sites they represent.

As an example, in Row 3 of Figure 1, the significant local maxima are indicated by red circles. As final results, the detected peaks were ranked according to their p-values. Of the 17993 significant peaks, the top 7918 had p-values that could not be distinguished from 0 because of the numerical accuracy of our Monte Carlo simulations. These peaks were ranked according to their SNR.

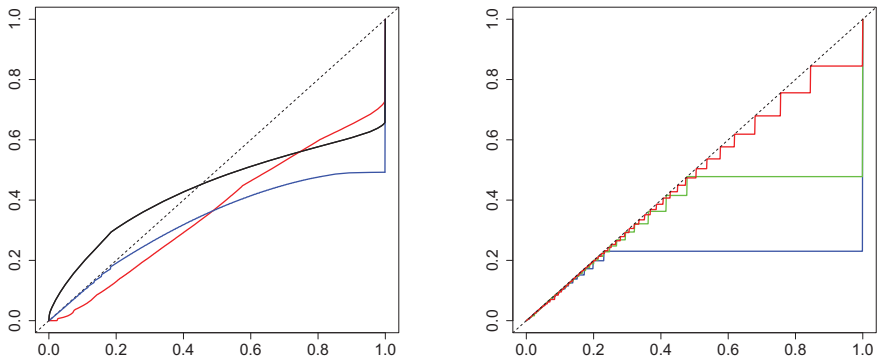
To give a sense of the amount of signal in the data and the validity of the procedure, Figure 5a compares the observed marginal distribution of p-values to the expected marginal distribution under the complete null hypothesis in the FoxA1 dataset. The observed marginal distribution of p-values (shown in black in Figure 5a) is given by the empirical distribution

$$\hat{G}(p) = \frac{1}{\tilde{m}} \sum_{t \in \tilde{T}} \mathbf{1}\left[p(t) \le p\right], \qquad 0 \le p \le 1, \tag{8}$$

where $\tilde{T}$ is the set of $\tilde{m}$ locations of the local maxima of $\widetilde{\mathrm{IP}}(t)$, with p-values given by (7). The marginal distribution under the complete null hypothesis is estimated in two different ways, one purely empirical and one more theoretical.

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press



a b

Figure 5: Marginal distribution of p-values in the FoxA1 dataset: (a) observed (black) and estimated under the global null hypotesis empirically (red) and theoretically (blue); (b) specific null distributions for $\lambda = \lambda_{\mathrm{L}} = 0.00097$ (red), $\lambda = 0.00135$ (blue), and $\lambda = 0.00256$ (green).

The empirical estimate (shown in red in Figure 5a) was obtained by running the entire analysis on the Control sample as if it was the IP, that is, searching for peaks in the Control sample using the same Control sample for estimating the background. The obtained null distribution of p-values lies below the diagonal as required for validity, and it exhibits a high frequency of p-values equal to 1, corresponding to peaks with only one tag in them.

The theoretical estimate (shown in blue in Figure 5a), was obtained as follows. Recall that for a smoothed stationary Poisson sequence $\tilde{X}(t;\lambda)$ with constant rate λ, the distribution of the height of a local maximum at $t \in \tilde{T}$ is given by (6). Analogous to (7), define the corresponding null p-value as $p_0(t) = F\left(\tilde{X}(t);\lambda\right)$ for $t \in \tilde{T}$. Its distribution $G_0(p;\lambda) = P(p_0(t) \le p)$ for any t is given by

$$G_0(p;\lambda) = \begin{cases} 1, & F(u_1;\lambda) \le p \\ F(u_k;\lambda), & F(u_k;\lambda) \le p < F(u_{k-1};\lambda), \quad k = 2,3,\ldots \end{cases} \tag{9}$$

where u_k, $k = 1,2,\ldots$, are the discrete values taken by the smoothed process $\tilde{X}(t;\lambda)$ at the local maxima. Note that $G_0(p;\lambda)$ is independent of t for $t \in \tilde{T}$ because of stationarity. In the ChIP-Seq problem, we approximate the null distribution of the p-value at $t \in \tilde{T}$ by the null distribution $G_0(p;\hat{\lambda}_0^+(t))$ corresponding to a stationary process with constant rate $\lambda = \hat{\lambda}_0(t)$, which depends on t only through the value of λ. Since each of the observed p-values in (8) corresponds to a different background rate $\hat{\lambda}_0(t)$, the estimated marginal distribution under the global null hypothesis is given by the mixture distribution

$$\hat{G}_0(p) = \frac{1}{\tilde{m}} \sum_{t \in \tilde{T}} G_0\left(p;\hat{\lambda}_0^+(t)\right), \qquad 0 \le p \le 1. \tag{10}$$

Referring back to Figure 5a, the observed distribution is always above the null distribution, and the large derivative at zero indicates the presence of a strong signal, which explains the large number of significant peaks found. Note that the null distribution is not uniform but stochastically larger. To better understand the mixture (10), Figure 5b shows three examples

COBRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper133

of the individual null distributions (9). All are discrete and stochastically larger than the continuous uniform distribution. For small λ, the most common p-value is 1 as most local maxima take the smallest possible value u_1, equal to the mode of the kernel $w(t)$, obtained when there is an isolated count of 1 in a neighborhood of zeros. This explains the large jump at 1 in panel (a). As λ gets larger, the distribution becomes closer to the continuous uniform distribution.

3 Comparison to other methods

As a reference, we compared our method to MACS and cisGenome on both the FoxA1 and GABP datasets. While the FoxA1 dataset was used in the development of MACS and our method, the GABP dataset was not used in the development of any of the three methods, providing an impartial test of performance. The methods were compared by a motif analysis and in terms of their mutual agreement.

On the FoxA1 dataset, the MACS and cisGenome software were applied using the default values. MACS returned a list of 13639 regions, which were ranked by p-value. In the case of cisGenome, binding regions were ranked by their FDR estimate and those with an FDR estimate of less than 0.05 were retained, yielding 7945 significant regions. In both cases, the locations of the peaks were identified by the reported summit within each region. This makes the results comparable to ours, since our method produces peak locations, rather than regions.

On the GABP dataset, all three methods were applied using the default values. Our method, thresholding at an FDR level of 0.01, produced 4072 significant peaks. MACS produced a list of 13828 peaks, while cisGenome produced a list of 4275 peaks with an FDR estimate of 0.05 or below.

3.1 Motif analysis

As biological validation, a motif analysis was performed where, for each peak declared significant, the number of motifs related to the appropriate transcription factor was counted within 100 bp and 400 bp of the estimated peak location. The distance of 400 bp approximately corresponds to the spatial spread of the measurements belonging to a binding site, as determined by the estimate in Figure 2a. Table 2 shows the average number of motifs and the proportion of peaks with at least one motif within those distances for the top 7945 peaks found by all methods in the FoxA1 dataset, and the top 4072 peaks found by all methods in the GABP dataset. In the FoxA1 dataset, the number 7945 is the number of peaks in the smallest of the three lists, in this case cisGenome. Taking the same number of top peaks in each list makes the averages and proportions in the table comparable, as the peak lists are ordered and the various methods use different criteria for their list cut-offs. In the GABP dataset, the number 4072 is again the number of peaks in the smallest of the three lists, in this case our method. The table indicates that our method, labeled "STEM+Regr" for simplicity, shows a similar but slightly better performance, particularly in the 400 bp range.

3.2 Peak overlap and discrepancies

To help explain the previous results, Table 3 shows the percentage of peaks from the top 7945 from each method in the FoxA1 dataset, or the top 4072 from each method in the GABP dataset, that were also found by each of the other methods within a distance of 100 bp and 400 bp. The matrices in the table are not symmetric because the correspondence between peaks is

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Table 2: Motif analysis comparing the performance of the proposed method against MACS and cisGenome on two different datasets. Results are for the top 7945 peaks in all methods for the FoxA1 dataset, and the top 4072 peaks in all methods for the GABP dataset.

		Average number of motifs within		Proportion with at least one motif within	
Dataset	Method	100 bp	400 bp	100 bp	400 bp
FoxA1	STEM+Regr	0.886	1.826	0.609	0.828
	MACS	0.888	1.812	0.609	0.827
	cisGenome	0.862	1.769	0.593	0.812
GABP	STEM+Regr	0.836	1.652	0.556	0.779
	MACS	0.832	1.650	0.557	0.776
	cisGenome	0.827	1.597	0.546	0.757

Table 3: Number of peaks from the methods listed in the columns that were also found by the methods listed in the rows within a distance of 100 bp and 400 bp. Results are for the top 7945 peaks from all methods in the FoxA1 dataset, and the top 4072 peaks from all methods in the GABP dataset.

		% found within 100 bp			% found within 400 bp		
Dataset	Method	STEM+Regr	MACS	cisGenome	STEM+Regr	MACS	cisGenome
FoxA1	STEM+Regr	100	80.3	76.1	100	84.2	79.2
	MACS	78.8	100	83.8	78.9	100	84.4
	cisGenome	74.0	83.8	100	74.1	84.3	100
GABP	STEM+Regr	100	90.9	81.8	100	94.3	85.2
	MACS	90.9	100	84.3	91.2	100	85.2
	cisGenome	81.8	84.3	100	82.3	85.2	100

not one-to-one; peaks found by one method may be represented by two or more peaks found by another method. The table shows that there is a fair amount of overlap (about 74% to 94%) between the methods, our method aligning better with MACS than with cisGenome.

To better understand the discrepancies, Figure 6 shows two examples of genomic segments from the GABP dataset after alignment. The left panel shows one of the 52 out of the top 4072 peaks in the list produced by our method that were not found among the 13828 peaks produced by MACS nor among the top 4275 peaks produced by cisGenome. Our method detected a secondary peak within 567 bp of a major peak (Row 3), in a binding region that was counted as a single region by both MACS and cisGenome. All the other peaks in this group of 52 were found to be secondary peaks or sometimes tertiary peaks. Of these, 48 were more than 400 bp away from its closest neighbor, with distances up to 1467 bp and averaging 604 bp. This indicates that these secondary peaks, not distinguished by the other methods, may be separate binding sites. The ability to resolve these peaks is a consequence of our method searching for binding sites, rather than binding regions.

The right panel of Figure 6 shows one of the 821 of the 13828 peaks produced by MACS that were also found among the list of 4275 produced by cisGenome but not among the top 4072 produced by our method. This peak was not called significant by our method because its

Harvard University Biostatistics Working Paper Series
Research Archive

http://biostats.bepress.com/harvardbiostat/paper133

associated p-value was not low enough (Row 6). This is because the peak height is low (notice the difference in vertical scale in Row 3 relative to the left panel), while the estimated local background rate is high (Row 4). Thus the SNR (Row 5) is lower than other peaks in the genome, pushing this one down the ranking list. In fact, the global regression model (4) for this dataset gave all the weight in the coefficients to the 1 Kb window and none to the larger windows or the GC coefficients, so the local background estimate at the peak (Row 4) directly reflects the high activity in the Control in the neighborhood of the peak (Row 1). Other peaks in this group of 821 are similar. This example illustrates the importance of the estimation of the local background rate in the analysis.

4 Discussion

4.1 Methodological considerations

We have presented a method for detection of peaks in ChIP-Seq data based on the STEM algorithm of SGA with promising results. The applicability of SGA to ChIP-Seq data relied on the common assumption that the signal peaks, represented by a mean function, are unimodal and have the same shape up to an amplitude scaling factor. The adaptation to ChIP-Seq data required two main modifications: 1) estimation of the local background rate; 2) use of Monte Carlo simulations to compute p-values.

From a methodological point of view, estimation of the background rate $\lambda_0(t)$ is arguably the most crucial step in the analysis, as the inference for a particular local maximum is highly dependent on the background rate at that location. In this paper, we have focused on the inference aspects of detecting peaks with a spatial structure via the STEM algorithm. Estimation of the local background rate (the particle "Regr" in the acronym "STEM+Regr") is not part of the original STEM algorithm, but is necessary for the analysis of ChIP-Seq data because the noise process is not stationary. This conceptual separation is helpful in that the background estimation method could be replaced by a different method if desired, without affecting the general implementation of the STEM algorithm for peak detection.

Statistical methods for estimating the variable rate $\lambda(t)$ in dynamic Poisson models or non-homogeneous Poisson sequences have been developed in other contexts. Bayesian methods (West et al., 1985; Harvey and Durbin, 1986; Bolstad, 1995) are computationally intensive, estimating $\lambda(t)$ at each location t based on the estimates at locations $1, 2, \ldots, t-1$. This is computationally infeasible for long genomic sequences as in ChIP-Seq data. Other methods require either repeated realizations (Arkin and Leenis, 2000) or a more specific structure of the process (Zhao and Xie, 1996; Helmers et al., 2003), which are not available in ChIP-Seq data.

In this paper, we have proposed a simple solution to local background estimation based on multiple linear regression. This method has the ability to easily incorporate the Control sample and other covariates such as the local GC content. In addition, estimating the regression coefficients from the data automatically adjusts for sequencing depth and the relative weighting between the various window sizes. Given this relative weighting, the method adaptively estimates the local background at each location. In this sense, the regression model solves the normalization problem and gives a partial answer to the question of how slowly $\lambda_0(t)$ varies with t. In the FoxA1 dataset, the three windows of length 1 Kb, 10 Kb and 100 Kb were given equal weight by the regression, while in the GABP dataset all the weight was given to the 1 Kb window. Because the GABP dataset is richer in number of reads, the regression method automatically accounts for it and allows estimation of the background rate at a smaller spatial scale.

Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

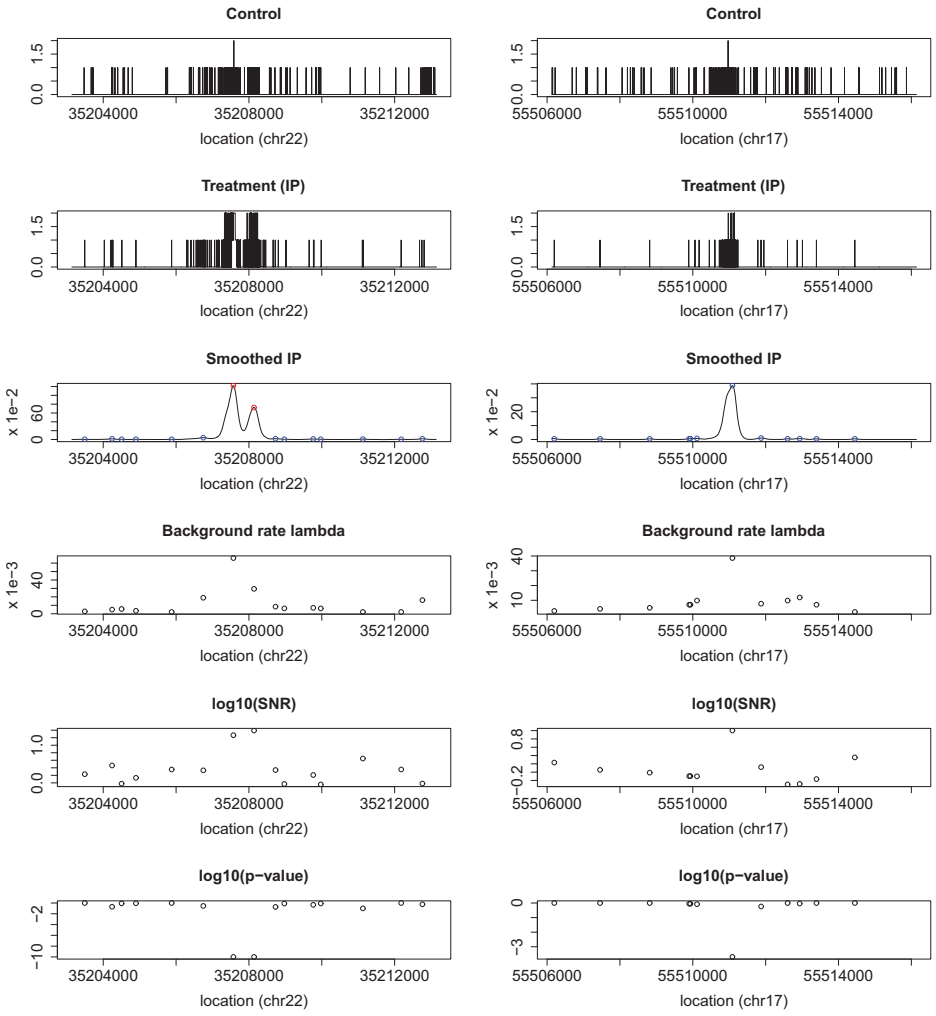


Figure 6: Left: A fragment of the aligned GABP data featuring a secondary peak called by our method but not MACS or cisGenome. Right: A fragment of the aligned GABP data featuring a peak called by MACS and cisGenome but not by our method. The variables plotted are the same as in Figure 1. Notice the difference in vertical scales between the left and right panels.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Often in ChIP-Seq data a Control sample is unavailable. In such case, the regression model (4) could have the 10 Kb and 100 Kb averages from the IP sample itself as predictors instead of those from the Control, with the 1 Kb window not included in the model. This would allow estimation of the background from the neighborhood of each peak, albeit with some positive bias. The GC content contribution would remain. Fortunately, the positive bias would make the inference more conservative, affecting the detection power more than its validity.

In the comparison with MACS and cisGenome, it was observed that the STEM algorithm performs competitively in terms of nearby motifs. In the datasets analyzed, all three methods found many of the same strong peaks. However, our method found secondary and terciary peaks near other strong peaks that were not distinguished by the other methods, and may possibly be separate binding sites. This is a result of our method searching for localized binding sites rather than binding regions of arbitrary size.

On the other hand, our method did not call significant other peaks that were called by the other methods. These peaks were not strong enough when compared to their corresponding background estimate at that location, at least according to the background estimation method used here. It is possible that a different background estimation method would have caused these peaks to be called significant. In fact, the other methods did because they had different assumptions about what represents a strong peak.

In this paper, we have attempted to frame the ChIP-Seq problem as a formal multiple testing problem. Assuming a Poisson model, we searched for peaks in the IP sample whose binding rate is higher than the Control at that location and higher than a minimally interesting rate. The significance results and FDR levels may be trusted as far those hypotheses are concerned. Moreover, because of the inherent spatial spread of the binding peak shape up to 400 bp and the spatial smoothing applied to the data, the detection procedure implictly defines a detected peak to be a true positive if it is within that distance of its true location. This helps explain the relatively low FDR cut-off of 0.01 used in the analysis. However the spatial spread also limits the spatial precision with which the detected peak locations can be estimated. The biological validity of the results is limited by the validity of the modeling assumptions. Particularly difficult is the background estimation, for which no good model exists to date. Because of its importance, background estimation is where future research in ChIP-Seq analysis should focus its attention.

4.2 Computational considerations

In addition to the modelling considerations mentioned above, the final ranking of the detected peaks depends on the numerical accuracy with which p-values are computed. In the Monte Carlo simulations for computing the distribution of the heights of local maxima, the estimation is more accurate for high values of λ, as these produce more observations. In the simulations, we set the simulation length to be at least 10^5, or as long as needed in order to obtain at least 100 Poisson counts. The latter condition was necessary for very low Poisson rates, but cannot be consider sufficient. The random variability was amelliorated by applying B-spline smoothing across λ in order to obtain the table in Figure 4b.

In the analysis results, we observed that a large number of detected peaks had a p-value of zero, meaning that the Monte Carlo simulation did not have enough numerical resolution to distinguish between their p-values. These peaks were ranked sub-optimally by SNR. More accurate calculation of p-values could be achieved with longer Monte Carlo simulations or by more sophisticated simulation techniques, such as Importance Sampling.

Computational complexity is also important in ChIP-Seq analysis because of the large

Collection of Biostatistics Research Archive

Hosted by The Berkeley Electronic Press

amount of data to be processed. The methods in this paper were implemented in `R` to ease their development and sharing among researchers, but at the expense of computational speed. The main computational bottleneck of our method is kernel smoothing, taking about $6 \sim 8$ hours to run over the entire genome on a Dell Power Edge R710 server with CPU speed 2.67 GHz, 48 GB of memory and a Linux CentOS 5.5 operating system. All the other processing steps together take about another hour. Kernel smoothing is mathematically simple, yet unfortunately inefficient in `R` for very long sequences. Computing time for kernel smoothing increases linearly with the kernel and the sequence size. In our implementation, the data was divided into subgroups of tags no more than 10^4 bp apart, trading off the length of the groups and their number. Computational time was also reduced by reducing the length of the kernel by multiplying it by a quartic biweight function of smaller support. In comparison to the other methods, cisGenome is fast because it is precompiled in `C`, while MACS is implemented in Python. Even though we implemented programming tricks for speed gains in `R`, such as run length encoding, in the future the ideas proposed here could be made computationally competitive by implementing them in `C`.

At this stage, we do not intend that the method proposed in this paper is viewed as a monolithic direct competitor to the already existing methods for analyzing ChIP-Seq data, but rather as a suggestion of how rich inferential theory for multiple testing in spatial domains, such the one described in SGA, can inform a more accurate detection of peaks, at least from a statistical point of view.

A software package implementing the methods in this paper are available at:
`http://www.biostat.jhsph.edu/~ajaffe/research.html`.

A Alignment details

A.1 Raw data

The FoxA1 raw data consists of a table of about 3.9 million rows for the IP sample and a table of about 5.2 million rows for the Control sample. Each row corresponds to a mapped tag of length 35 bp and contains the begininng and end genomic addresses for the tag and an indicator of whether the tag belongs to the forward (+) or reverse (−) DNA strand. We define the location of a tag to be given by its beginning address, corresponding to the lower address for the forward (+) tags and the higher address for the reverse (−) tags. The GABP data set, containing about 7.8 million tags in the IP sample and about 17.4 million tags in the Control sample, was converted to the same format before processing. Genomic locations not listed in the table were assumed to have an associated tag count of zero. Duplicate tags were considered measurement artifacts and were removed from the analysis.

In order to be counted together, the tags from the two strands need to be aligned with each other. We followed an alignment method similar to that in MACS, shifting all tags by the same amount in the 3' direction of the tag sequence toward the most likely binding site: forward (+) tags toward higher genomic addresses and reverse (−) tags toward lower genomic addresses. Once shifted, tags coinciding at the same location are counted together. The result of this process is a table of genomic locations, each with an associated tag count. This aligned data is used as the input for peak detection, described in Section 2.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper133

A.2 Estimation of the tag shift and peak shape

As in MACS, we estimate the size of the shift from the tag count distributions corresponding to a set of strong and easily detectable peaks, as described below. We performed the shift estimation on Chromosome 1 because of its likelihood to contain enough such strong peaks, but other long chromosomes could be used instead. As part of the process, the shift estimation also allows us to estimate the distribution of shifted tags counts around a peak. This peak shape, normalized to unit sum, is used later as a smoothing kernel in the STEM algorithm for peak detection. The estimation proceeds as follows.

Algorithm 2 (Estimation of shift size and peak shape)**.**

1. Temporarily shift all tags ((+) forward and (−) back) by a tentative shift amount (default 100 bp). This produces a table of genomic locations, each with an associated tag count.

2. Perform peak detection on the count data from the previous step and select a set of strong peaks (details given below). Let $t_1, \ldots, t_N$ be their locations.

3. Set a window size W (an odd number, default 2001 bp). The distribution of the forward tags is a vector of length W whose i-th entry is equal to the average number of forward tags at a constant distance $(W+1)/2-i$ from the peak, that is, at locations $t_j-(W+1)/2+i$, $j = 1, \ldots, N$. Repeat for the reverse tags.

4. Fit a spline to the distribution of forward tags and record its mode. Repeat for the reverse tags. The estimated shift is half the distance between the two modes, rounded to the nearest integer.

5. To estimate the peak shape, shift the original forward and reverse distributions by the estimated shift, symmetrize the joint distribution by averaging both the forward and reverse tag distributions and their mirror images with respect to the center of the window, and fit a spline.

The peak detection step (Step 2) above need not be exact. Since the tag distribution is evaluated in a window around the strong peaks, it is enough that the true location of those peaks is contained somewhere near the center of that window. To achieve this, we apply the first half of the STEM algorithm, as follows.

2a. Set a tentative unimodal symmetric kernel (default Gaussian with standard deviation 50) and perform kernel smoothing on the count data from Step 1. (Implementation details given in Section 2.3).

2b. Find the local maxima of the smoothed count sequence. (Implementation details given in Section 2.3).

2c. Select the N highest local maxima (default 1000).

At the end of this process, the data consists of a long sequence of genomic addresses and associated counts 0, 1 or 2, ready for peak detection analysis. The maximal count of 2 is a result of the elimination of duplicates from the original list of tags. Because binding rates are generally low, truncation at 2 does not greatly affect the Poisson model used thereafter.

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

References

Bradford L. Arkin and Lawrence M. Leenis. Nonparametric estimation of the cumulative intensity function for a nonhomogeneous Poisson process from overlapping realizations. *Management Science*, 46(7):989–998, 2000.

A. Barski and K. Zhao. Genomic location analysis by ChIP-Seq. *Journal of Cellular Biochemistry*, 107:11–18, 2009.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc B*, 57(1):289–300, 1995.

W. M. Bolstad. The multiprocess dynamic Poisson model. *J. Am. Statist. Assoc.*, 90(429): 227–232, 1995.

A. Fejes, G. Robertson, M. Bilenky, R. Varhol, M. Bainbridge, and S. Jones. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, 24(15):1720–1730, 2008.

A. C. Harvey and J. Durbin. The effects of seat belt legislation on British road casualties: A case study in structural time series modelling. *J. Royal Statist. Soc.*, 149:187–227, 1986.

Roelof Helmers, I. Wayan Mangku, and Ričardas Zitikis. Consistent estimation of the intensity function of a cyclic Poisson process. *J. Multivar. Anal.*, 84(1):19–39, 2003.

Hongkai Ji, Hui Jiang, Wenxiu Ma, David S. Johnson, Richard M. Myers, and Wing H. Wong. An integrated software system for analyzing ChIP-chip and ChIP-Seq data. *Nature Biotechnology*, 26(11):1293–1300, 2008.

D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316:1497–1502, 2007.

T. S. Mikkelsen, M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T. K. Kim, R. P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E. S. Lander, and B. E. Bernstein. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448:553–560, 2007.

Peter J. Park. ChIP-seq: advantages and challenges of a maturing technology. *Nature Rev. Genet.*, 10:669–680, 2009.

William K. Pratt. *Digital image processing*. Wiley, New York, 1991.

Armin Schwartzman, Yulia Gavrilov, and Robert J. Adler. Multiple Testing of Local Maxima for Detection of Unimodal Peaks in 1D. Working paper, Harvard University, `http://www.bepress.com/harvardbiostat/paper131/`, 2011.

Marvin Simon. *Digital communication techniques: signal design and detection*. Prentice Hall, Englewood Cliffs, NJ, 1995.

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

DIGITAL COMMONS
A BEPRESS REPOSITORY
Harvard University Biostatistics
Research Archive

A. Valouev, D. S. Johnson, A. Sundquist, C. Medina, E. Anton, S. Batzglou, R. M. Myers, and A. Sidow. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature Methods*, 5:829–834, 2008.

M. West, P. J. Harrison, and H. S. Migon. Dynamic generalized linear models and Bayesian forecasting (with discussion). *J. Am. Statist. Assoc.*, 80:73–96, 1985.

Yong Zhang, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nussbaum, Richard M. Myers, Myles Brown, Wei Li, and X. Shirley Liu. Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137, 2008.

M. Zhao and M. Xie. On maximum likelihood estimation for a general non-homogeneous Poisson process. *Scand. J. Statist.*, 23(4):597–607, 1996.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press