ChIP-chip: Data, Model, and Analysis

Ming Zheng[1], Leah O. Barrera[2], Bing Ren[3] and Ying Nian Wu[1,4] *
Department of Statistics, UCLA[1].
Ludwig Institute for Cancer Research, UCSD[2].
Department Of Cellular and Molecular Medicine, UCSD School Of Medicine[3].

Abstract: ChIP-chip (or ChIP-on-chip) is a technology for isolation and identification of genomic sites occupied by specific DNA binding proteins in living cells. The ChIP-chip data can be obtained over the whole genome by tiling arrays, where a peak in the signal is generally observed at a protein binding site. In this paper, we describe and present a probability theory for modelling ChIP-chip data. We then propose a model-based computational method for locating and testing peaks for the purpose of identifying potential protein binding sites.

Keywords: Genome, Mpeak, Peak detection, Protein binding sites, Sonication, Truncated triangle shape model.

1 Introduction

ChIP-chip (or ChIP-on-chip) [22, 18, 9, 21, 4, 17, 11, 16, 13, 14, 5, 6, 20], also known as genome-wide location analysis, is a technology for isolating genomic sites occupied by specific DNA binding proteins in living cells. This strategy may be used to annotate functional elements, such as promoters, enhancers, repressor elements, and insulators, in genomes by mapping the locations of protein markers associated with these sites.

In the term "ChIP-chip," "ChIP" refers to "chromatin immunoprecipitation," which is a method for isolating DNA fragments that are bound by specific DNA binding proteins.

*1. Department of Statistics, UCLA, 8125 Math Sciences Bldg, Los Angeles, CA 90095-1554, USA. 2. Ludwig Institute for Cancer Research, UCSD, 9500 Gilman Drive, La Jolla, CA 92093-0653, USA. 3. Department Of Cellular and Molecular Medicine, UCSD School Of Medicine, 9500 Gilman Drive, La Jolla, CA 92093-0653, USA. 4. To whom correspondence should be addressed. Email: ywu@stat.ucla.edu

"Chip" refers to the DNA microarray technology [19] for measuring the concentrations of these DNA fragments. The DNA microarray probes can tile the whole genome, so the ChIP-chip data can be obtained over the whole genome in the form of a one-dimensional signal, where a peak in the signal is generally present at a protein binding site. Therefore, the protein binding sites can be located by detecting the peaks in the signal. For the purpose of peak detection, it is desirable to develop mathematical models for the ChIP-chip data.

The model must be probabilistic in nature, mainly because the chromatin immunoprecipitation process involves cutting the long genomic sequences into small DNA fragments by sonication, and this process is a stochastic one. In this paper, we derive the functional forms of the ChIP-chip data under probabilistic assumptions about this process.

After studying the probability model of ChIP-chip data, we shall describe a model-based computational method for locating and testing the peaks for the purpose of identifying potential protein binding sites. We then illustrate our method using the recent data obtained by Kim et al. [12]. They used ChIP with microarrays, tiling all non-repetitive sequence of the human genome with 50 bp probes at 100 bp resolution, to obtain a high-resolution map of active promoters in one human cell type in steady state.

A software called Mpeak has been developed based on the method proposed in this article. The software (including the visual C++ source code) is free to download.

2 ChIP-chip data

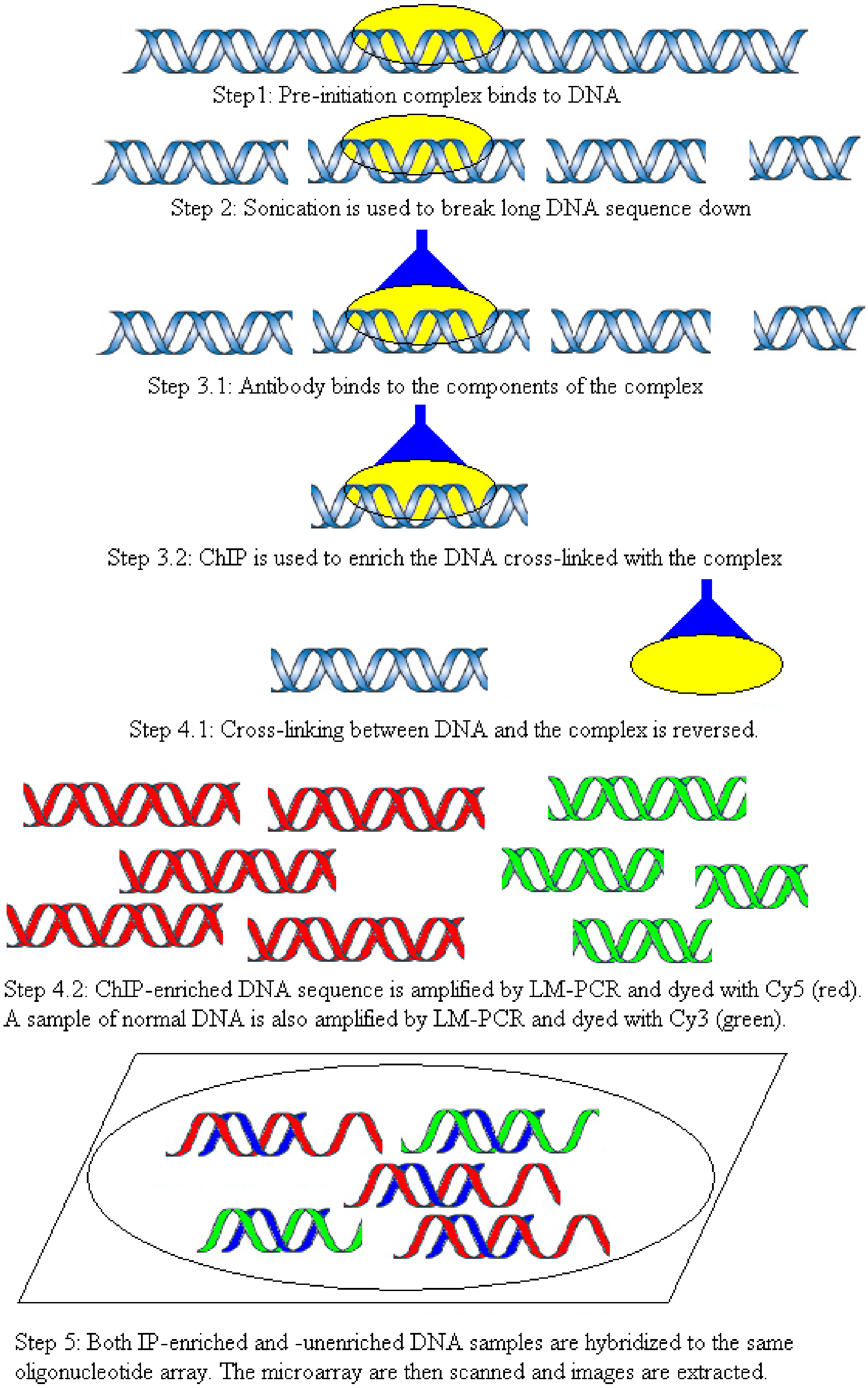
This section gives a description of the ChIP-chip process. Readers who are more interested in mathematical modelling can jump directly to Section 3, and then come back to this section for details.

The ChIP-chip process is shown in Figure 1.

Step 1: Let proteins bind to DNA: bound transcription factors and other DNA-associated proteins are cross-linked to DNA with formaldehyde.

Step 2: Chop the DNA sequences into small fragments: Sonication is used to break genomic DNA to small DNA fragments while the transcription factors are still bound to DNA. Therefore, among all the chopped DNA fragments, some are bound by proteins, and the rest are not.

Step 3: Isolate the DNA fragments bound by proteins (or the DNA fragments containing protein binding sites) by chromatin immunoprecipitation (ChIP). For instance, in Kim



Step1: Pre-initiation complex binds to DNA

Step 2: Sonication is used to break long DNA sequence down

Step 3.1: Antibody binds to the components of the complex

Step 3.2: ChIP is used to enrich the DNA cross-linked with the complex

Step 4.1: Cross-linking between DNA and the complex is reversed.

Step 4.2: ChIP-enriched DNA sequence is amplified by LM-PCR and dyed with Cy5 (red). A sample of normal DNA is also amplified by LM-PCR and dyed with Cy3 (green).

Step 5: Both IP-enriched and -unenriched DNA samples are hybridized to the same oligonucleotide array. The microarray are then scanned and images are extracted.

Figure 1: Illustration of ChIP-chip method.

et al. [12], an antibody specifically recognizing a component of the pre-initiation complex, i.e. the TAF1 subunit of the general transcription factor IID (TFIID) is added and used to immunoprecipitate DNA fragments corresponding to the promoter regions bound by TAF1.

Step 4: Cross-linking between DNA and protein is reversed and DNA is released, amplified by LM-PCR (here, we ligate linkers to DNA fragments that allow us to amplify them all at the same time using the same set of primers by PCR) and labelled with a fluorescent dye (Cy5). At the same time, a sample of DNA which is not enriched by the above immunoprecipitation process are also amplified by LM-PCR and labelled with another fluorescent dye (Cy3).

Step 5: both IP-enriched and -unenriched DNA pools of labelled DNA are hybridized to the same high-density oligonucleotide arrays (chip). The microarray is then scanned and two images, corresponding to Cy5 (TAF1 IP) and Cy3 (control), respectively, are extracted.

Intensity-dependent Loess [10] can be used to normalize the resulting signal values for both images, and median filtering (window size = 3 probes) can be applied to smooth the log(Cy5/Cy3) data.

Kim et al. [12] used this method to analyze the active promoters in human genome. They used antibodies specially recognizing components of the transcription pre-initiation complex to obtain a high-resolution map of active promoters in human genome. Using this approach, they were able to annotate transcriptional start sites and discover novel genes. The data analyzed in this paper come from their experiment, but the algorithm is generally applicable to other ChIP-chip experiments where peak finding can be used to localize the binding sites of a transcription factor of interest.

3 Probability Modelling

In this section, we derive probability models for ChIP-chip data. The probabilities calculated for one random genome sequence manifest themselves as frequencies among the large number of genome sequences in an experiment. The derivations of the formula are elementary and non-rigorous so that they are easy to follow for interested biologists.

3.1 ChIP process

Genome and binding sites: The protein binding sites (such as promoters) on the genome can be modelled as a set of points on the real line. Let's denote the locations of these binding sites by their coordinates $B_1, B_2, ..., B_M$. The total number M of binding sites and their coordinates are unknown, and need to be inferred from the ChIP-chip data.

Protein binding: In the ChIP-chip experiment, the proteins are bound to the binding sites. For a genome sequence, let p_m be the probability that the binding site m is bound by a protein. The binding at different binding sites are assumed to be independent of each other.

Sonication: The sonication process chops the genome sequences into short DNA fragments. Each fragment is an interval on the real line. For a genome sequence, the set of cut points are randomly distributed.

A common probability model is the Poisson point process model, which has the following assumptions: 1) the probability that a cut point occurs in a small interval $(x, x + \Delta x)$ is $\lambda(x)\Delta x$, where $\lambda(x)$ is the intensity function measuring how dense the cut points are around x. $1/\lambda(x)$ can be considered the expected length of the intervals between two consecutive cut points around x. 2) For non-overlapping intervals, what is happening in one interval is independent of what is happening in the other intervals.

The Poisson model can be considered the first order approximation to reality. It captures the marginal information about the density of cut points. The interactions between cut points are not modelled.

Immunoprecipitation: For each protein bound to a binding site, the probability that it is bound by the antibody is α. For a DNA fragment to be immunoprecipitated, it must contain at least one binding site that is bound by protein, which must in turn be bound by the antibody. We call such a binding site a "good binding site." Clearly, the probability that B_m is a good binding site is $p_m\alpha = q_m$. A DNA fragment that contains at least one good binding site is called a "good fragment."

Tiling array of probes: At each location x, the array signal measured by a probe at x is denoted by $Y(x) = \log(\text{Cy5}/\text{Cy3})$. It measures the relative abundance of good fragments that contain x.

3.2 Probability calculation

Consider a random genome sequence. The ChIP process produces from this genome sequence a collection of non-overlapping good fragments. These good fragments only cover part of the whole genome.

For any location x, let $p(x)$ be the probability that x is covered by a good fragment. In the following, we shall calculate $p(x)$ under various scenarios.

For x to be covered by a good fragment, a necessary and sufficient condition is that there is no cut point between x and at least one good binding site.

One binding site scenario: Let's first consider the simplest scenario where there is only one binding site at the origin of the real line. Then

$$\begin{aligned} p(x) &= Pr(0 \text{ is a good binding site and} \\ &\quad \text{no cut point between 0 and x}) \\ &= q \times Pr(\text{no cut in (0,x)}) \end{aligned} \tag{1}$$

where q is the probability that 0 is a good binding site, i.e., it is bound by a protein, which is in turned bound by the antibody. Without loss of generality, let's assume that $x > 0$.

To compute Pr(no cut $\in (0, x)$), we can divide the interval $(0, x)$ into a large number of small bins, $(0, \Delta x)$, $(\Delta x, 2\Delta x)$, ... ,$(i\Delta x, (i + 1)\Delta x)$, ..., $((n - 1)\Delta x, n\Delta x)$, where $\Delta x = x/n$. Let $x_i = i\Delta x$. According to the Poisson assumption,

$$\begin{aligned} \Pr \quad & (\text{no cut} \in (0, x)) \\ = \quad & \prod_{i=1}^{n} \Pr(\text{no cut} \in ((i-1)\Delta x, i\Delta x)) \\ = \quad & \prod_{i=1}^{n} (1 - \lambda(x_i)\Delta x). \end{aligned} \tag{2}$$

Taking log on both sides,

$$\begin{aligned} \log \quad & \Pr(\text{no cut} \in (0, x)) \\ = \quad & \sum_{i=1}^{n} \log(1 - \lambda(x_i)\Delta x) \\ = \quad & \sum_{i=1}^{n} [-\lambda(x_i)\Delta x + o(\Delta x)] \\ \to \quad & -\int_0^x \lambda(s) ds, \quad \text{as } n \to \infty, \end{aligned} \tag{3}$$

where Taylor expansion gives us $\log(1-\lambda(x_i)\Delta x) = -\lambda(x_i)\Delta x + o(\Delta x)$, with $o(\Delta x)$ being a term that decreases to 0 faster than $1/n$ as $n \to \infty$. Thus

$$\log p(x) = \log q - \int_0^x \lambda(s)ds, \quad \text{for } x > 0. \tag{4}$$

If we assume $\lambda(x) = a$ for $x > 0$, then

$$\log p(x) = c - ax, \quad \text{for } x > 0,$$

where $c = \log q$. Similarly for $x \leq 0$, if we assume $\lambda(x) = b$, then

$$\log p(x) = c + bx, \quad \text{for } x \leq 0.$$

We can combine the above two equations into one equation,

$$\log p(x) = c - b[-x]^+ - a[x]^+, \tag{5}$$

where $[x]^+ = x$ if $x > 0$, and $[x]^+ = 0$ otherwise. From the above equation, it is easy to see that $\log p(x)$ has a triangle shape with peak at 0.

Equation (5) is the basis for our model-based peak detection method. However, this model assumes that there is only one binding site. For real data, the above model is true only around a local neighborhood of a binding site, where the effects of other binding sites can be neglected. In the following, we shall study the situation of more than one binding site, in order to understand how different binding sites affect each other.

We can also view this problem from survival analysis perspective. Here, for simplicity, we still assume that there is only one good binding site at the origin 0 and we consider some point $x > 0$. We can regard the good binding site as the time point that a patient receives treatment. Then, the event that there is no cut between 0 and x is equivalent to the event that the patient survives time point x. The probability is simply the survival function of the patient, and the $\lambda()$ function is the hazard rate.

Two binding site scenario: Suppose there are two binding sites B_1 and B_2. Let's assume that $B_1 \leq B_2$. Let q_1 and q_2 be the probabilities that they are good binding sites respectively. For $x \in (B_1, B_2)$, $p(x)$ is influenced by both B_1 and B_2.

$$\begin{aligned} p(x) &= \Pr(B_1 \text{ is good and no cut} \in (B_1, x) \text{ or} \\ &\quad B_2 \text{ is good and no cut} \in (x, B_2)) \\ &= q_1 \Pr(\text{no cut} \in (B_1, x)) \\ &\quad + q_2 \Pr(\text{no cut} \in (x, B_2)) \\ &\quad - q_1 q_2 \Pr(\text{no cut} \in (B_1, B_2)) \\ &= q_1 \exp\{-\int_{B_1}^{x} \lambda(s) ds\} \\ &\quad + q_2 \exp\{-\int_{x}^{B_2} \lambda(s) ds\} \\ &\quad - q_1 q_2 \exp\{-\int_{B_1}^{B_2} \lambda(s) ds\}, \qquad (6) \end{aligned}$$

where the last step follows the same logic as equations (2) and (3).

If B_1 and B_2 are far away from each other, and if x is close to B_1, then the last two terms in equation (6) can be neglected, and we will obtain an approximated equation that is in the same form as (4) in the one binding site scenario.

General scenario: Now we are ready to derive the formula for general scenario, where there are M binding sites $B_1, ..., B_M$. For notational convenience, we also add $B_0 = -\infty$, and $B_{M+1} = \infty$, with $q_0 = q_{M+1} = 0$. For $x \in (B_m, B_{m+1})$,

$$\begin{aligned} p(x) &= \Pr(\text{no cut between } x \text{ and the nearest} \\ &\quad \text{good binding site to the left} \\ &\quad \text{or no cut between } x \text{ and the nearest} \\ &\quad \text{good binding site to the right}) \\ &= p_L(x) + p_R(x) - p_L(x) p_R(x), \qquad (7) \end{aligned}$$

where

$$\begin{aligned} & p_L(x) \\ &= \Pr(\text{no cut between } x \text{ and the nearest} \\ &\quad \text{good binding site to the left}) \\ &= \sum_{i=m}^{0} \Pr(\text{the nearest good binding site} \qquad (8) \\ &\quad \text{to the left is } B_i \text{ and no cut} \in (B_i, x)) \\ &= \sum_{i=m}^{0} \left[\prod_{j=i+1}^{m} (1 - q_j) \right] q_i \exp\{-\int_{B_i}^{x} \lambda(s) ds\}. \end{aligned}$$

$$\begin{aligned}
&p_R(x) \\
&= \Pr(\text{no cut between } x \text{ and the nearest}r \\
&\quad \text{good binding site to the right}) \qquad (9)\\
&= \sum_{i=m+1}^{M+1} \left[\prod_{j=i-1}^{m+1} (1-q_j) \right] q_i \exp\{-\int_x^{B_i} \lambda(s)ds\}.
\end{aligned}$$

With equations (8) and (9), $p(x)$ can be computed according to equation (7).

From the above analysis, we can see that the triangle shape fits the data only within a local range around a true binding site. So in our data analysis, we shall fit a truncated triangle shape model whose range is adaptively determined.

3.3 Chip measurement

The "chip" part of the ChIP-chip technique is intended to measure $\log p(x)$. In particular, the Cy5 measures the abundance of DNA fragments in the IP-enriched DNA pool, and Cy3 measures the abundance of DNA fragments in the unenriched DNA pool. For a DNA fragment containing probe x, the hybridization strength, i.e., the probability that it will be hybridized by the probe x, can depend on x. By computing $Y(x) = \log(\text{Cy5}/\text{Cy3})$, this dependence is cancelled out.

There has been previous work [23, 8] on modelling the chip data. We shall simply assume that the errors are distributed with constant marginal variance.

4 Model fitting and peak detection

4.1 Fit truncated triangle shape model

In order to make inference about $M, B_1, ..., B_M$ from the observed signal $Y(x)$, ideally one may adopt a Bayesian modelling and inference framework, using the model derived for the general scenario in the previous subsection, by making assumptions on the smoothness of $\lambda(s)$, as well as the form of measurement noise. However, the computation can be too expensive given the length of the genome. Instead, we propose to locally fit the following approximated model:

$$\begin{aligned}
Y(x) &= \log(\text{Cy}_5/\text{Cy}_3) \\
&= c - b[B-x]^+ - a[x-B]^+ + \epsilon(x), \qquad (10)
\end{aligned}$$

which is a triangle shape model, where $\epsilon(x)$ is assumed to be a Gaussian process with constant marginal variance.

We may fit this triangle shape model on the data around each probe to see if the model fits the data. Let us use x_0 to denote the position of this probe. We look at a window around x_0. Let's denote the probes on the left of x_0 by $(x_{-L}, ..., x_{-1})$, and the probes on the right of x_0 by $(x_1, ..., x_R)$. Let the signals measured by these probes be $(y_{-L}, ..., y_{-1}, y_0, y_1, ..., y_R)$. We then fit the following multiple regression model

$$y_i = c - b[x_0 - x_i]^+ - a[x_i - x_0]^+ + \epsilon_i, -L \leq i \leq R, \tag{11}$$

by least squares estimate. To be more specific, let

$$Y = (y_i)_{i=-L}^R, \quad X = (1, -[x_0 - x_i]^+, -[x_i - x_0]^+)_{i=-L}^R,$$

where Y is the column vector composed of y_i for $i = -L, ..., R$, and X is the 3-column matrix. Then the least squares estimates of the parameters are

$$\begin{aligned} (\hat{c}, \hat{b}, \hat{a})' &= (X'X)^{-1}(X'Y), \\ \hat{\sigma}^2 &= \|Y'Y - Y'X(X'X)^{-1}X'Y\|^2/(R + L + 1). \end{aligned} \tag{12}$$

4.2 Peak finding

With the ability to fit the triangle shape model, we propose the following peak finding algorithm.

1 Identify all the local maximum probes in the data. A probe is a local maximum probe if its signal is greater than all the signals within k bp away (k is a parameter that is pre-specified and the default number is 200).

2 As a starting point, pick the probe with the largest signal among all the local maximum probes.

3 At the current probe x, fit the triangle shape model as described above, for all combinations of (L, R), where both L and R are chosen within a range from the smallest allowable value to the largest allowable value (these two values are pre-specified, and the default numbers are 300 bp and 1500 bp respectively). Then choose the (L, R) that gives us the smallest residual variance $\hat{\sigma}^2$. We call $(x - L, x + R)$ the range of this probe x, and $\hat{\sigma}^2$ the residual of x.

4 Repeat the above model fitting procedure for the neighbors of this current local maximum probe. For each neighboring probe x, obtain its range and residual as described in step 2. Then among the current local maximum probe and its neighbors, choose the probe with the smallest residual. We mark this probe as a potential binding site.

5 For any smaller local maximum in the range of this best fitted triangle, we compare the fit of the triangle just identified and that of the triangle centered at the smaller local maximum. If the difference between the two fitted values at the smaller local maximum are smaller than a threshold (which is a factor times the SD of the residuals of the best fitted triangle, and the default factor is 1.5), then the smaller one is said to be inhibited by the larger one and marked as non-peak.

6 Among all the local maximum probes still not marked, choose the local maximum probe with the largest signal. Then go back to step 3. Stop the algorithm if all the local maxima are marked, or below a threshold (a pre-specified value, whose default value is the mean plus 2.5 SD of the raw data).

Despite the large amount of computation involved, the algorithm takes less than one minute to analyze a genome long sequence on a regular PC.

4.3 Peak testing

For a potential binding site x, suppose the triangle shape model fitted at x covers n probes. Let $Y_1, Y_2, ..., Y_n$ be the signals of these n probes, which can be considered the signals caused by the potential binding site x. We want to test whether x is a real binding site.

We decide to use the following test statistic:

$$\bar{Y}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i.$$

A similar method to calculate P-value is proposed by Buck, Nobel, and Lieb ([7]), where background noise is assumed to be independent. Some properties of statistic of this kind are discussed elsewhere ([1]).

If $Y_1, ..., Y_n$ are not caused by a binding site, they should be pure noise, which can be modelled by a stationary process. This process is not independent white noise, because there are auto-correlations between nearby probes. We may assume that Y_i is correlated with its neighbors Y_j with $|P_j - P_i| \leq m$ (P_j and P_i are the genomic positions of Y_j and Y_i, respectively). Then

$$\begin{aligned}
& \mathrm{Var}(\bar{Y}_n) \\
= \; & \mathrm{Var}(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i) = \frac{1}{n} \sum_{i,j} \mathrm{Cov}(Y_i, Y_j) \\
= \; & \frac{1}{n} \sum_{|P_i - P_j| \leq m} \mathrm{Cov}(Y_i, Y_j) \\
\approx \; & \mathrm{Var}(Y_i)(1 + \sum_{|P_j - P_i| \leq m} \mathrm{Cov}(Y_i, Y_j)/\mathrm{Var}(Y_i)) \\
= \; & \gamma^2(1 + f),
\end{aligned}$$

where γ^2 is the marginal variance $\mathrm{Var}(Y_i)$, and f is the auto-correlation factor. Both can be estimated from the data.

Then we can obtain the p-value by comparing the observed $\bar{Y}_n$ with $N(0, \gamma^2(1 + f))$. The normal distribution can be justified by the central limit theorem. We can trim the insignificant peaks by thresholding the p-value.

5 Results on real data

We have applied our algorithm to real data obtained from human genome. The reader is referred to [12] for biological discoveries and validations. Figures below show some examples of model fitting. The plot on the top shows the observed signals. The plot in the middle shows the signals produced by the fitted triangle shape models. The plot on the bottom shows the probes that are considered the potential binding sites.

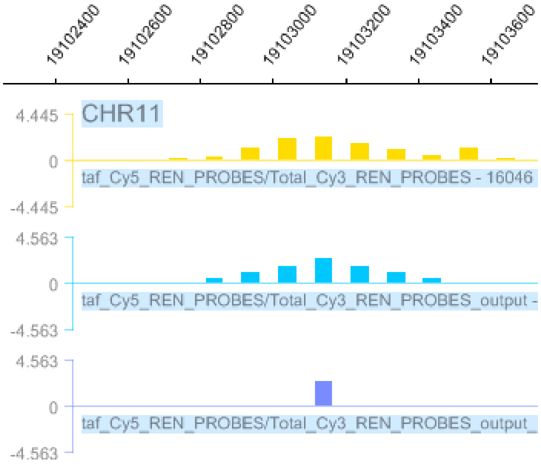


Figure 2: Top: original data. Middle: fitted data. Bottom: Peak position.

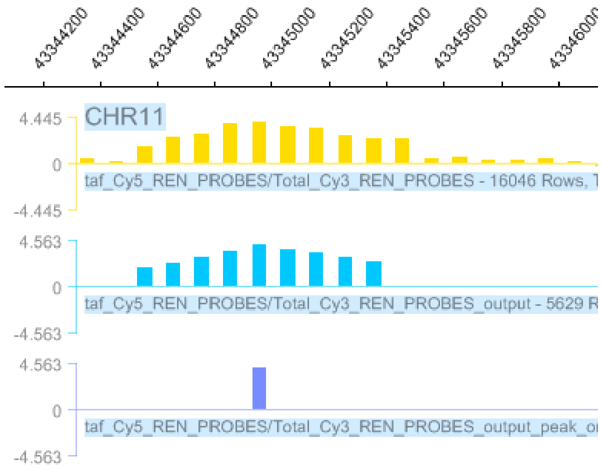


Figure 3: Top: original data. Middle: fitted data. Bottom: Peak position.

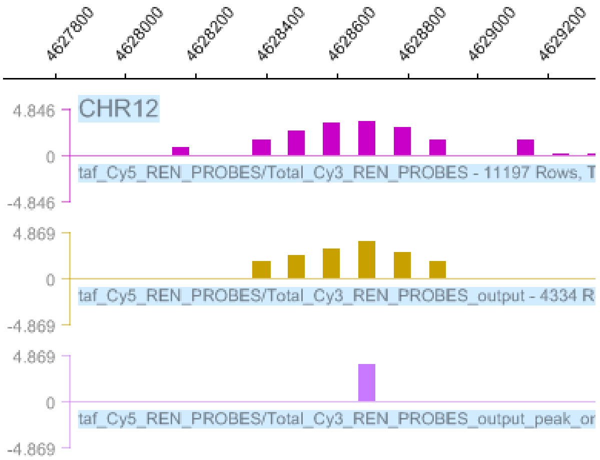


Figure 4: Top: original data. Middle: fitted data. Bottom: Peak position.

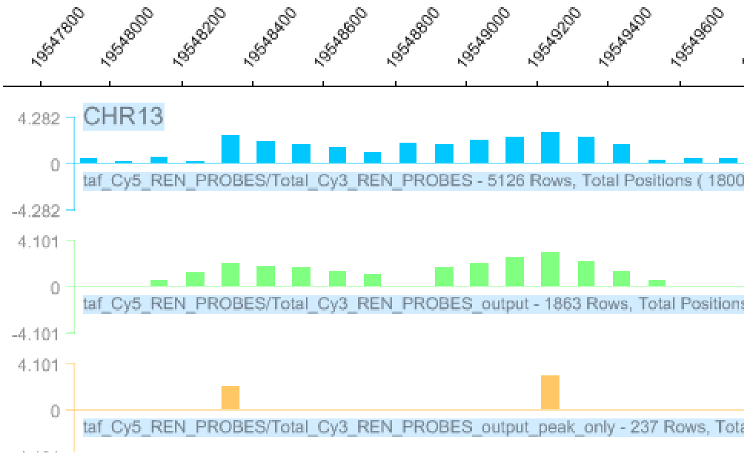


Figure 5: Top: original data. Middle: fitted data. Bottom: Peak position.

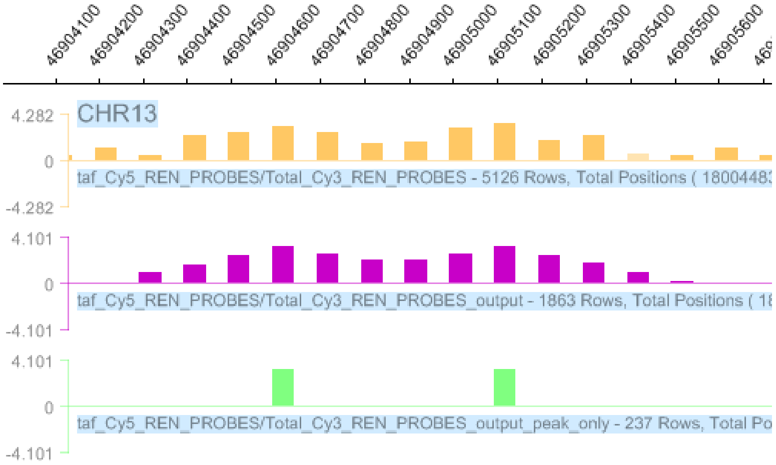


Figure 6: Top: original data. Middle: fitted data. Bottom: Peak position.

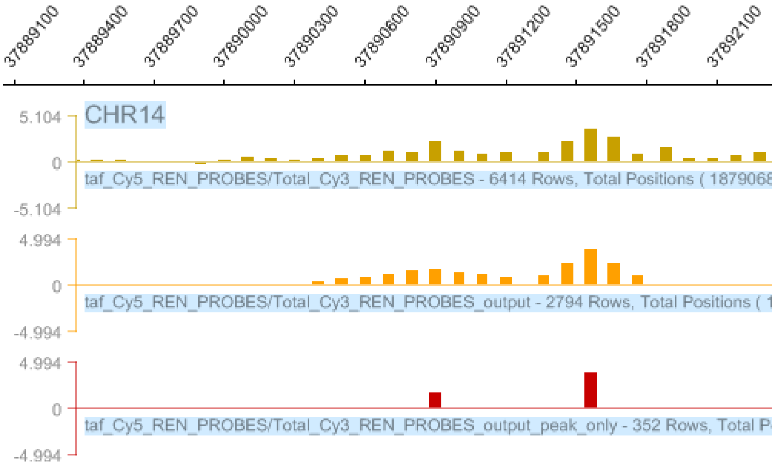


Figure 7: Top: original data. Middle: fitted data. Bottom: Peak position.

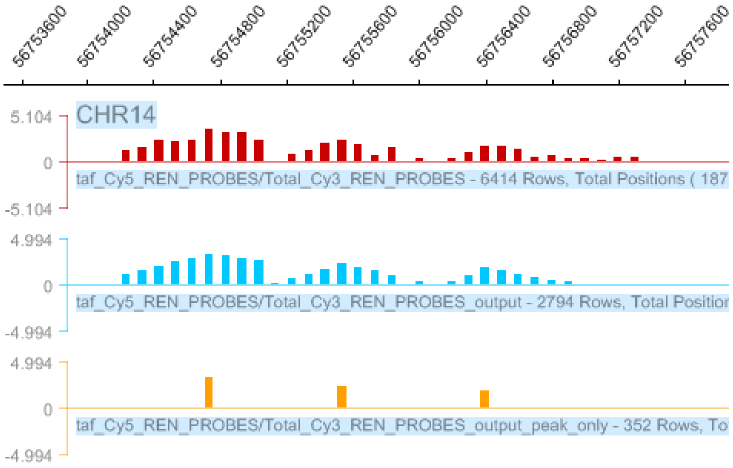


Figure 8: Top: original data. Middle: fitted data. Bottom: Peak position.

It appear that the triangle shape model provides reasonable fit to the observed data. But there are also cases where the model fitting is not very good. See Figures 9 and 10.

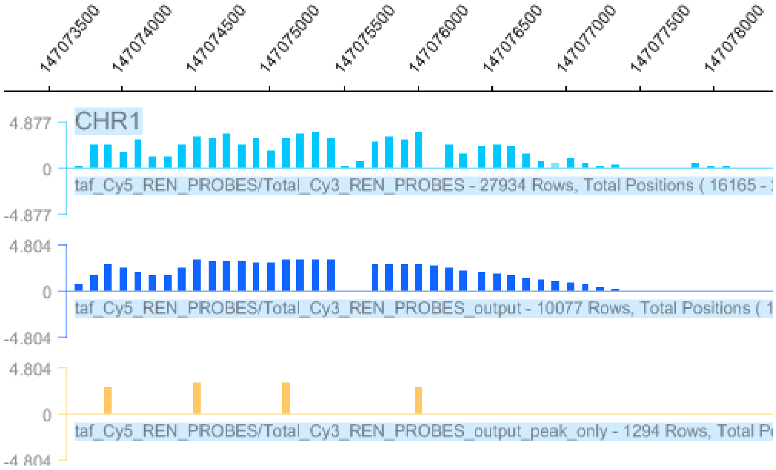


Figure 9: Top: original data. Middle: fitted data. Bottom: Peak position.

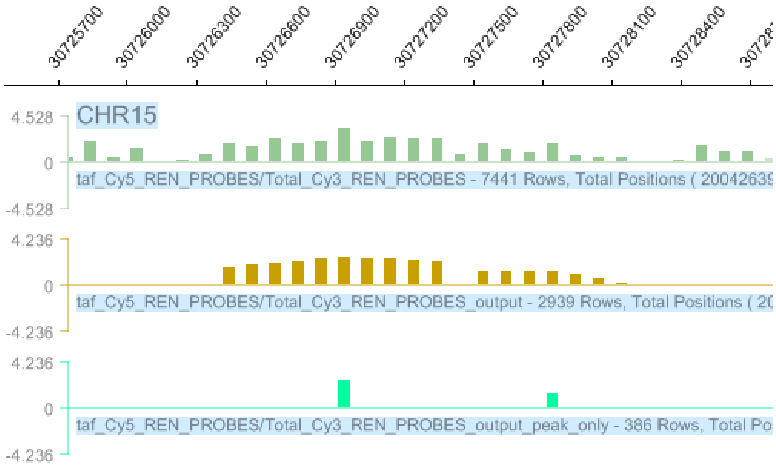


Figure 10: Top: original data. Middle: fitted data. Bottom: Peak position.

To assess the overall fitness of the triangle shape model, we compute the R^2 measure for every detected peak. Specifically, for a peak at x, let the range of the peak be $[x-L, x+R]$. We compute the overall variance of all the signals within this range,

$$\hat{\tau}^2 = \sum_{i=x-L}^{x+R} (Y_i - \hat{\mu})^2/(R+L+1),$$

where $\hat{\mu}$ is the average of all the $Y_i, i \in [x-L, x+R]$. Then $R^2 = 1 - \hat{\sigma}^2/\hat{\tau}^2$, where $\hat{\sigma}^2$ is the estimated residual variance of the triangle shape model, see equation (12). R^2 measures the percentage of the variance explained by the triangle shape. The larger R^2

is, the better the model fits the data. Figure 11 plots the histogram of the R^2 for all the detected peaks, indicating that the model fits the data reasonably well.

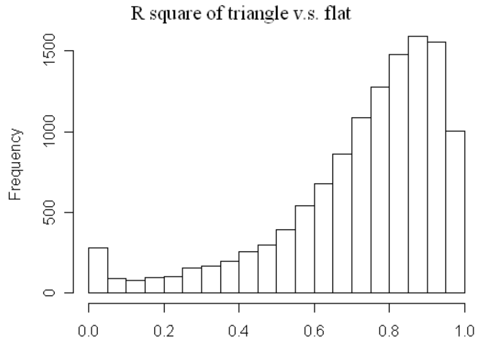


Figure 11: Histogram of R^2 measures for all the detected peaks.

Another indicator of goodness of fit is the log likelihood ratio score, which is $(R + L + 1)\log(\hat{\tau}^2/\hat{\sigma}^2)$. The triangle shape model has three more parameters than the model assuming a common mean (i.e. a flat top instead of a triangle shape), namely, the peak position and the two slopes of the triangle shape. Figure 12 plots the log-likelihood ratio scores of the detected peaks. Most of the log-likelihood ratio scores are much larger than 11.34, the 99% quantile of χ^2_3, suggesting that the triangle shape model provides considerable improvements over the flat top model.

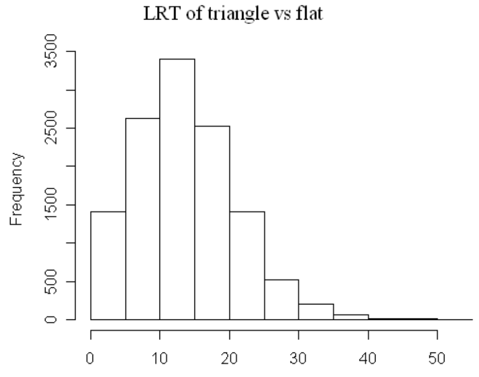


Figure 12: Histogram of log-likelihood ratios for all the detect peaks.

6 Justifications of truncated triangle model

The following are justifications for fitting the truncated triangle model.

1) Probabilistically, the exponential model is a local approximation to the underlying sonication process.

2) Functionally, the truncated triangle model is a local linear approximation to the functional form around the true binding site.

3) Empirically, the model provides reasonable fit to the data. We may not have enough data and computing power to fit more sophisticated non-linear models.

4) The signals produced by a binding site are not only characterized by a large value at the corresponding probe, but also by a roughly triangle shape caused by the binding site (figure 13). In order to rank and test the peaks in the data, it is advantageous to identify a neighborhood of probes whose signals may be caused by a biding site.

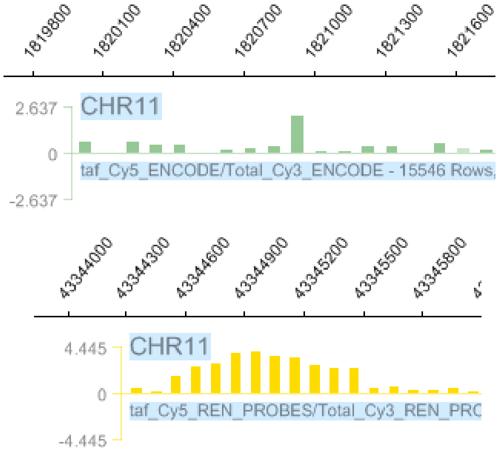


Figure 13: Illustration of necessity of model fitting: the single large signal in the first figure is likely to be noise, and the signals in the second figure is likely to indicate a true binding site.

5) Our model fits the range and two slopes to the data. This can be more adaptive than finding local maxima after smoothing the data with kernel function of fixed bandwidth (e.g., [7]). In some sense, our model fitting method corresponds to finding the kernel bandwidth adaptively.

7 Mpeak software

A software named Mpeak has been developed following our peak finding procedure and released to public. The software and the source code are free to download from www.stat.ucla.edu /$\sim$ zmdl /mpeak. To use the software, please just download the file, unzip it and follow the instruction. Currently, only windows-executable version is provided. The software is reasonably fast: $\sim$ 10 seconds for $\sim$ 400,000 probes on a regular PC (Celeron 2.4GHz, 512 Mb memory.)

8 Discussion

The contribution of this article lies in two aspects. 1) We provide a theoretical framework for modelling the ChIP-chip process, and we derive the functional forms of the ChIP-chip data under simple assumptions. 2) We develop a model-based algorithm for locating the positions and ranges of the peaks that are caused by potential binding sites. We also provide a method for testing the significance of the peaks compared to the background noise.

There are a number of issues that need to be addressed in further work.

1) The Poisson model is only a first order approximation to the spatial distribution of the cut points caused by the sonication process. Its validity needs to be carefully examined, in comparison with more sophisticated models, such as Markov point process models, which take into account the interactions between the cut points.

2) The chip process for measuring $\log p(x)$ needs to be modelled more carefully. In particular, the auto-correlation structures in the tiling array data need to be studied theoretically and empirically.

3) The model fitting algorithm needs improvement. Currently, the algorithm only seeks to explain the data around the potential binding sites. We need to develop efficient algorithms to fit the whole data set using the multiple-peak model. The model should also allow rare cases where probes near the binding sites do not hybridize efficiently and result in poor signals.

4) In computing the p-values, we do not take into account the fact that the peaks are selected by an optimization algorithm. As a result, the p-values can be smaller than they should be. We may use a simulation method to obtain more accurate p-values, although this can be time consuming. Currently, the p-values are only used as indications of the significance of the peaks, for the purpose of trimming insignificant peaks.

5) FDR controlling method proposed by Benjamini et al ([2], [3]) can be incorporated to take care of the multiple testing scenario in our data.

6) The model-based peak detection method can be extended to detecting more sophisticated shapes.

References

[1] Arias-Castro, E., Donoho, D.L. and Huo, X. (2003) Near-Optimal Detection of Geometric Objects by Fast Multiscale Methods. Stanford Statistics Department, Technical Report, 2003.

[2] Benjamini, Y., and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing, In: *Journal of the Royal Statistical Society*, **B** 57, No. 1, PP.: 289-300.

[3] Benjamini, Y., and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency, In: *The Annals of Statistics*, 2001, Vol. 29, No. 4, PP.: 1165-1188.

[4] Bernstein, B.E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D.K., Huebert, D.J., McMahon, S., Karlsson, E.K., KulbokasIII, E.J., Gingeras, T.R., Schreiber, S.L. and Lander, E.S. (2005) Genomic Maps and Comparative Analysis of Histone Modifications in Human and Mouse. *Cell*, 2005, Vol. 120, PP.: 169-181.

[5] Boguski, M.S. (2004) ENCODE and ChIP-chip in the genome era. *Genomics*, 2004, Vol. 83, PP.: 347-348.

[6] Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., Gifford, D.K., Melton, D.A., Jaenisch, R. and Young, R.A. (2005) Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells. *Cell*, 2005, Vol. 122, PP.: 947-956.

[7] Buck, M.J., Nobel, A.B. and Lieb, J.D. (2005) ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. *unpublished.*

[8] Callow, M.J., Dudoit, S., Gong, E.L., Speed, T.P., and Rubin, E.M. (2000) Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Research*, 2000, Vol. 10, PP.: 2022-2029.

[9] Cawley, S., Bekiranov, S., Ng, H.H., Kapranov, P., Sekinger, E.A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A.J., Wheeler, R., Wong, B., Drenkow, J., Yamanaka, M., Patel, S., Brubaker, S., Tammana, H., Helt, G., Struhl, K., and Gingeras, T.R. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, 2004, Vol. 116, PP.: 499-509.

[10] Dudoit, S., Yang, Y.h., Callow, M.J., and Speed, T.P. (2000) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical report.

[11] Hanlon, S.E. and Lieb, J.D. (2004) Progress and challenges in profiling the dynamics of chromatin and transcription factor binding with DNA microarrays. *Current Opinion in Genetics and Development*, 2004, Vol. 14, PP.: 697-705.

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

[13] Kirmizis, A. and Farnham, P.J. (2004) Genomic Approaches That Aid in the Identification of Transcription Factor Target Genes. *Experimental Biology and Medicine*, 2004, 229, PP.: 705-721.

[14] Kirmizis, A., Bartley, S.M., Kuzmichev, A., Margueron, R., Reinberg, D., Green, R. and Farnham, P.J. (2004) Silencing of human polycomb target genes is associated with methylation of histone H3 Lys 27. *Genes and Development*, 2004, Vol. 18, PP.: 1592-1605.

[15] Klein, J.P. and Moeschberger, M.L. (2003) Survival analysis. *Springer Press*, 2003, 2nd edition.

[16] Lee, C.K., Shibata, Y., Rao, B., Strahl, B.D. and Lieb, J.D. (2004) Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nature Genetics*, 2004, 36, PP.: 900-905.

[17] Li, W., Meyer, C.A. and Liu, X.S. (2005) A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics*, 2005, Vol. 21, PP.: i274-i282.

[18] Li, Z., Calcar, S.V., Qu, C., Cavenee, W.K., Zhang, M.Q., and Ren, B. (2003) A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. *PNAS*, 2003, Vol. 100, No. 14, PP.: 8164-8169.

[19] Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E.L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 1996, Vol. 14, PP.: 1675-1680.

[20] Martone, R., Euskirchen, G., Bertone, P., Hartman, S., Royce, T.E., Luscombe, N.M., Rinn, J.L., Nelson, F.K., Miller, P., Gerstein, M., Weissman, S. and Snyde, M. (2003) Distribution of NF-kappaB-binding sites across human chromosome 22. *PNAS*, 2003, Vol. 100, No. 21, PP.: 12247-12252.

[21] Pokholok, D.K., Harbison, C.T., Levine, S., Cole, M., Hannett, N.M., Lee, T.I., Bell, G.W., Walker, K., Rolfe, P.A., Herbolsheimer, E., Zeitlinger, J., Lewitter, F., Gifford, D.K. and Young, R.A. (2005) Genome-wide Map of Nucleosome Acetylation and Methylation in Yeast. *Cell*, 2005, Vol. 122, PP.: 517-527.

[22] Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T.L., Wilson, C.J., Bell, S.P., and Young, R.A. (2000) Genome-wide location and function of DNA binding proteins, *Science*, 2000, Vol. 290, PP.: 2306-2309.

[23] Schadt, E.E., Li, C., Su, C., and Wong, W.H. (2000) Analyzing high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry*, 2000, Vol. 80, PP.: 192-202.