DOI: 10.1111/j.1541-0420.2007.00768.x

ChIP-chip: Data, Model, and Analysis

Ming Zheng,[1] Leah O. Barrera,[2] Bing Ren,[3] and Ying Nian Wu[1,*]

[1]Department of Statistics, UCLA, 8125 Math Sciences Bldg, Los Angeles,
California 90095-1554, U.S.A.
[2]Ludwig Institute for Cancer Research, UCSD, 9500 Gilman Drive, La Jolla,
California 92093-0653, U.S.A.
[3]Department of Cellular and Molecular Medicine, UCSD School of Medicine, 9500 Gilman Drive,
La Jolla, California 92093-0653, U.S.A.
[*]*email:* ywu@stat.ucla.edu

Summary. 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 signals can be obtained over the whole genome by tiling arrays, where a peak shape is generally observed around a protein-binding site. In this article, we describe the ChIP-chip process and present a probability model for ChIP-chip data. We then propose a model-based method for recognizing the peak shapes for the purpose of detecting protein-binding sites. We also investigate the issue of bandwidth in nonparametric kernel smoothing method.

Key words: Genome; Peak detection; Protein binding sites; Sonication; Truncated triangle shape model.

1. Introduction

ChIP-chip, also known as ChIP-on-chip or genome-wide location analysis (e.g., Ren et al., 2000), is a technology for isolating genomic sites occupied by specific DNA-binding proteins in living cells. This technology can be used to annotate functional elements in genomes, such as promoters, enhancers, repressor elements, and insulators, by mapping the locations of protein markers associated with these sites.

In the term "ChIP-chip," "ChIP" stands for "chromatin immunoprecipitation," which is a technology for isolating DNA fragments that are bound by specific DNA-binding proteins. "Chip" refers to the DNA microarray technology (Lockhart et al. 1996) for measuring the concentrations of these DNA fragments. The DNA microarray probes can tile the whole genome, so that the ChIP-chip data can be obtained over the whole genome in the form of a one-dimensional series of signals, where a peak shape is generally present around a protein-binding site. Therefore, the protein-binding sites can be located by recognizing the peak shapes in the signals.

For the purpose of peak recognition, it is desirable to develop mathematical models for the ChIP-chip data. The model is probabilistic in nature, 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 article, we derive the functional forms of the ChIP-chip data under simple probabilistic assumptions about this process.

After studying the probability model of ChIP-chip data, we describe a model-based method for recognizing the peak shapes for the purpose of pinpointing protein-binding sites. We then illustrate our method using data obtained by Kim et al. (2005).

2. ChIP-chip Data

This section gives a description of the ChIP-chip process, which is illustrated 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 sequences into 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 by chromatin immunoprecipitation (ChIP). For instance, in Kim et al. (2005), an antibody specifically recognizing a component of the preinitiation complex, 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 ligation-mediated polymerase (LM-PCR) chain reaction and labeled with a fluorescent dye (Cy5). At the same time, a sample of DNA, which is not enriched by the above immunoprecipitation process, is also amplified by LM-PCR and labeled with another fluorescent dye (Cy3).

Step 5: Both IP-enriched and -unenriched DNA pools of labeled DNA are hybridized to the same high-density oligonucleotide arrays (chip). The microarray is then scanned and

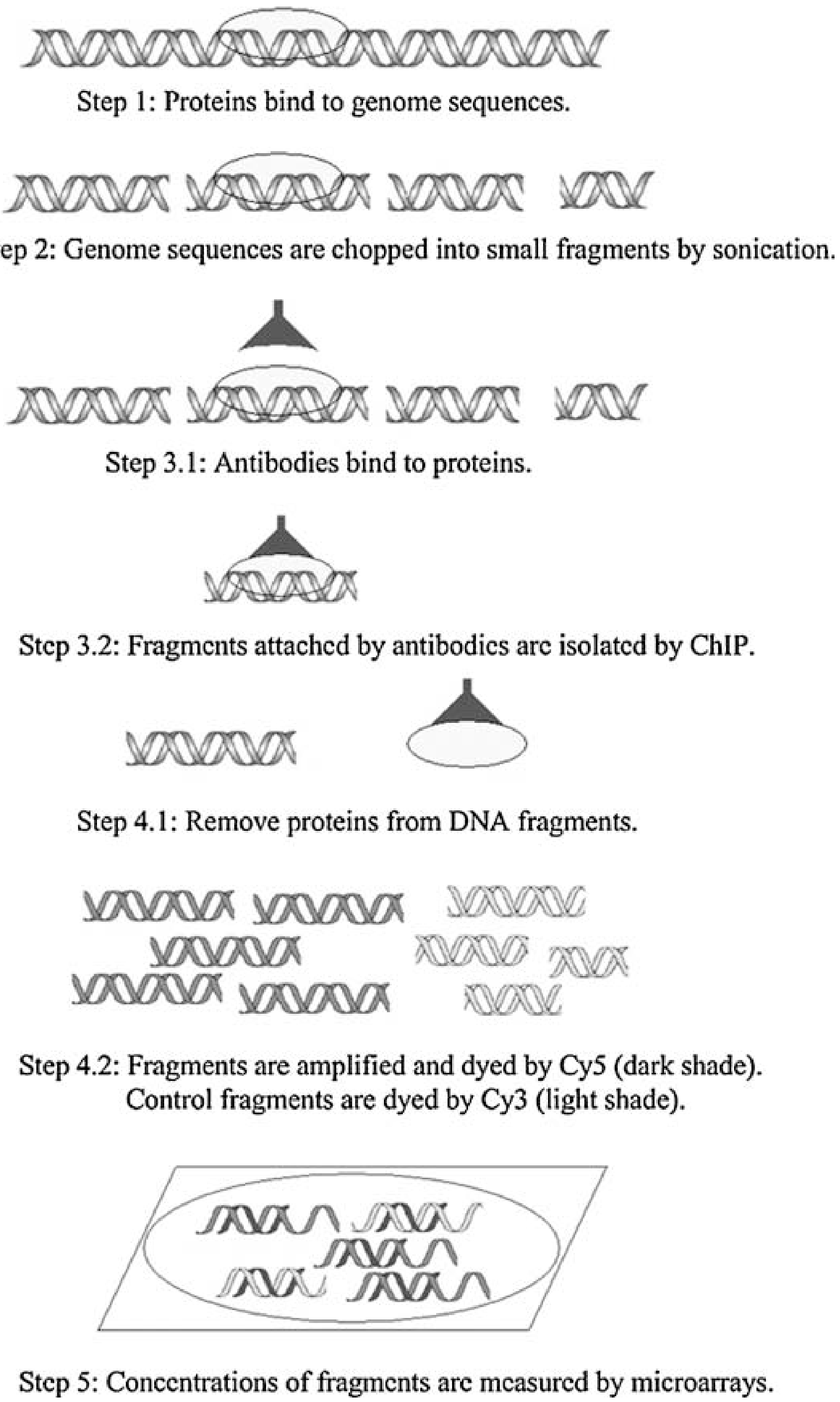
Step 1: Proteins bind to genome sequences.
Step 2: Genome sequences are chopped into small fragments by sonication.
Step 3.1: Antibodies bind to proteins.
Step 3.2: Fragments attached by antibodies are isolated by ChIP.
Step 4.1: Remove proteins from DNA fragments.
Step 4.2: Fragments are amplified and dyed by Cy5 (dark shade).
Control fragments are dyed by Cy3 (light shade).
Step 5: Concentrations of fragments are measured by microarrays.

Figure 1. Illustration of ChIP-chip process.

two images corresponding to Cy5 (TAF1 IP) and Cy3 (control), respectively, are extracted.

Intensity-dependent Loess (Dudoit et al., 2000) can be used to normalize the resulting signal values for both images. Median filtering (window size = 3 probes) can be applied to smooth the log(Cy5/Cy3) data.

3. Probability Modeling

In this section, we derive probability models for ChIP-chip data.

3.1 *ChIP Process*

Genome and binding sites: The protein-binding sites (such as promoters) on the genome can be idealized as a set of points on the real line. Let us denote the locations of these binding sites by their coordinates $B_1, B_2, \ldots, B_M$. The total number M of binding sites and their coordinates are unknown, and are 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 is 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 is randomly distributed.

A simple probability model is the Poisson point process, 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 nonoverlapping intervals, what is happening in one interval is independent of what is happening in the other interval.

Immunoprecipitation: For each protein bound to a binding site, the probability that it is recognized and 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 the protein, which must in turn be recognized and bound by the antibody. We call such a binding site a "good binding site." 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 $Y(x) = \log(\mathrm{Cy5}/\mathrm{Cy3})$. It measures the relative abundance of ChIP fragments that contain x.

The actual binding sites are generally several base pairs (bp) long, and the probes can be as long as 50 bp. Here we mathematically idealize them as dimensionless points on the real line for simplicity.

3.2 *Probability Model*

Consider a random genome sequence. The ChIP process produces from this genome sequence a collection of nonoverlapping 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 experiment, there are a large number of genome sequences, and $p(x)$ manifests itself as the concentration of good fragments covering x. So $\log p(x)$ can be considered the theoretical prediction of the signal value measured by probe x. In the following, we calculate $p(x)$ under various scenarios. In order to make this subsection easy to follow by interested biologists, we add some nonrigorous elementary steps in the derivations.

A key observation is: 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 us first consider the simplest scenario where there is only one binding site at the origin of the real line. Then,

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

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

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

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

The last step follows the Taylor expansion: $\log(1-\lambda(x_i)\ \Delta x) = -\lambda(x_i)\ \Delta x + o(\Delta x)$, with $o(\Delta x)$ being a term that decreases to 0 faster than $1/n$ as $n \to \infty$. Thus,

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

If we assume $\lambda(x) = a$ for $x > 0$, then $\log p(x) = c - ax$, for $x > 0$,where $c = \log q$. Similarly for $x \le 0$, if we assume $\lambda(x) = b$, then $\log\ p(x) = c + bx$, for $x \le 0$. We can combine the two equations for $x > 0$ and $x \le 0$ into one equation,

$$\log p(x) = c - b\,[-x]^+ - a\,[x]^+, \tag{2}$$

where $[x]^+ = x$ if $x > 0$, and $[x]^+ = 0$ otherwise.

Equation (2) has a triangle shape peaked at 0, and is the basis for our model-based peak recognition 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 from other binding sites can be neglected. In the following, we study the situation where there is more than one binding site, in order to understand how different binding sites affect each other.

Two binding sites scenario: Suppose there are two binding sites B_1 and B_2. Let us assume that $B_1 < B_2$. 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.

$$\begin{aligned} p(x) &= \Pr(B_1 \text{ is good and no cut} \in (B_1,\ x) \text{ or } B_2 \text{ is} \\ &\quad \text{good and no cut} \in (x, B_2)) \\ &= q_1 \exp\left\{-\int_{B_1}^{x} \lambda(s)ds\right\} + q_2 \exp\left\{-\int_{x}^{B_2} \lambda(s)ds\right\} \\ &\quad - q_1 q_2 \exp\left\{-\int_{B_1}^{B_2} \lambda(s)ds\right\}, \end{aligned} \tag{3}$$

where the last step follows the same logic as equation (1).

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 (3) can be neglected, and we will obtain an approximated equation that is in the same form as (2) in the one binding site scenario.

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

$$\begin{aligned} p(x) &= \Pr(\text{no cut} \in (x, \text{ nearest good binding site to the left}) \\ &\quad \text{or no cut} \in (x, \text{ nearest good binding site to the right})) \\ &= p_L(x) + p_R(x) - p_L(x)p_R(x), \end{aligned} \tag{4}$$

where

$$\begin{aligned} p_L(x) &= \Pr(\text{no cut} \in (x, \text{ nearest good binding site to the left})) \\ &= \sum_{i=0}^{m} \Pr(\text{nearest good binding site to the left is } B_i \text{ and} \\ &\quad \text{no cut} \in (B_i, x)) \\ &= \sum_{i=0}^{m} \left[\prod_{j=i+1}^{m} (1-q_j) \right] q_i \exp\left\{ -\int_{B_i}^{x} \lambda(s)\,ds \right\}. \end{aligned} \tag{5}$$

$$\begin{aligned} p_R(x) &= \Pr(\text{no cut} \in (x, \text{ nearest good binding site to the right})) \\ &= \sum_{i=m+1}^{M+1} \left[\prod_{j=m+1}^{i-1} (1-q_j) \right] q_i \exp\left\{ -\int_{x}^{B_i} \lambda(s)\,ds \right\}. \end{aligned} \tag{6}$$

With equations (5) and (6), $p(x)$ can be calculated according to equation (4).

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 the data analysis, we fit a truncated triangle shape model whose range is adaptively determined.

3.3 *Chip Measurement*

The "chip" step of the ChIP-chip process measures $\log p(x)$. 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 calculating $Y(x)=\log(\text{Cy5}/\text{Cy3})$, this dependence is cancelled out. We simply assume that the observational errors are additive and follow a stationary Gaussian process.

4. Model Fitting and Peak Recognition

The previous section shows that a binding site causes an approximately truncated triangle shape for the signals of the probes around this binding site. In this section, we propose a model-based method to recognize these shapes. After finding these truncated triangle shapes, including their positions and ranges, we can pool the probe signals within the range of each identified shape to test against the background noise hypothesis, to decide whether these signals are caused by a true binding site.

4.1 *Fit Truncated Triangle Shape Model*

The truncated triangle shape model is attempted to fit the data around each probe, and the positions and ranges of the shapes are identified by the best-fitted models.

Let x_0 denote the genomic coordinate of a probe. We fit the model within a window around x_0. Let L be the number of probes to the left of x_0 within the window. Let R be the number of probes to the right of x_0 within the window. Let us denote the genomic coordinates of the probes to the left of x_0 by $(x_{-L}, \ldots, x_{-1})$, and the coordinates of the probes to the right of x_0 by $(x_1, \ldots, x_R)$. Let the signals measured by these probes be $(y_{-L}, \ldots, y_{-1}, y_0, y_1, \ldots, y_R)$. We then fit the following multiple regression model,

$$y_i = c - b[x_0 - x_i]^+ - a[x_i - x_0]^+ + \epsilon_i, \quad -L \le i \le R, \tag{7}$$

where $a \ge 0$ and $b \ge 0$. We fit this model by constrained least squares method. Let $Y = (y_i)_{i=-L}^{R}$, and $X = (1, -[x_0 - x_i]^+, -[x_i - x_0]^+)_{i=-L}^{R}$. Then the least squares estimates of the coefficients are $(\hat{c}, \tilde{b}, \tilde{a})' = (X'X)^{-1}X'Y$. To satisfy the positivity constraints, we let $\hat{a} = [\tilde{a}]^+$ and $\hat{b} = [\tilde{b}]^+$. Because of DNA packaging and interactions with histones etc., there is reason to believe that the chopping rates around different binding sites may be different during the sonication step. Therefore, we assume that each peak has its own slopes a and b.

Let $\hat{Y} = X(\hat{c}, \hat{b}, \hat{a})'$. We calculate the residual variance $\hat{\sigma}^2 = \|Y - \hat{Y}\|^2/(L + R + 1 - d)$, where d is the number of regression coefficients. If both L and R are nonzero, then $d = 3$. If $L = 0$ or $R = 0$, then $d = 2$.

The residual variance $\hat{\sigma}^2$ is used for identifying the peak positions as well as the ranges L and R. It is not used for testing the significance of the peaks. Specifically, model (7) is correct under the following two assumptions: (1) x_0 is a true binding site, and (2) $\lambda(s)$ is constant within $[-L, 0)$ and $(0, R]$, respectively. If either assumption is incorrect, then model (7) is incorrect, and the residual variance $\hat{\sigma}^2$ will include the contribution from model bias. Therefore, a true binding site can be detected by the local minimum of the fitted $\hat{\sigma}^2$.

To be more specific, for any x_0 and L, R, let the signal $y_i = f(x_i) + \epsilon_i$. $f(x)$ is a truncated triangle shape peaked at x_0 if and only if assumptions (1) and (2) hold. If x_0 is not a true binding site, then $f(x)$ will not be a truncated triangle shape peaked at x_0. Instead, it will be a triangle peaked at a binding site other than x_0. Let $f = (f(x_i))_{i=-L}^{R}$ and $\epsilon = (\epsilon_i)_{i=-L}^{R}$. We can write $Y = f + \epsilon$. Let $H = X(X'X)^{-1}X'$ be the projection matrix, and let $\hat{Y} = HY, \hat{f} = Hf$, and $\hat{\epsilon} = H\epsilon$ be, respectively, the projections of Y, f and ϵ onto the space spanned by X. Then $E\|Y - \hat{Y}\|^2 = \|f - \hat{f}\|^2 + E\|\epsilon - \hat{\epsilon}\|^2$, because $E[\epsilon] = 0$. If assumptions (1) and (2) hold, then $f(x_i) = c - b\,[x_0 - x_i]^+ - a\,[x_i - x_0]^+$, so $\|f - \hat{f}\|^2 = 0$. If we shift x_0 from the true binding site while keeping L and R fixed, then $\|f - \hat{f}\|^2 > 0$. Assuming that ϵ_i come from a stationary process, and assuming that the probes are equally spaced, then $E\|\epsilon - \hat{\epsilon}\|^2$ remains unchanged under the shifting, because X remains the same. Therefore, $E\|Y - \hat{Y}\|^2$ or $E(\hat{\sigma}^2)$ is a local minimum relative to the shifting operation if assumptions (1) and (2) hold. This fact does not depend on the assumption that ϵ_i are uncorrelated. Therefore, we may use the residual variance $\hat{\sigma}^2$ to identify the locations of the binding sites.

We also use the residual variance $\hat{\sigma}^2$ to determine the ranges L and R of the truncated triangle shape. If ϵ_i is uncorrelated with constant marginal variance σ^2, then under assumption (1), $E(\hat{\sigma}^2) = \sigma^2$ for any L and R that satisfy assumption (2). If L or R is too large for assumption (2) to be true because of the effects from nearby binding sites, then $E(\hat{\sigma}^2) > \sigma^2$. In practice, we choose L and R that give us minimum $\hat{\sigma}^2$ among all the allowable combinations of L and R. This is a conservative choice. L and R determine the range of a fitted triangle shape, so that we can pool the signals within this range and use their average to test against the background hypothesis. For a peak shape caused by a true binding site, the conservative choice of L and R already enables us to include the strong signals around the binding site. Even though the conservative choice of L and R may fail to include the relatively weak signals of the probes that are near the two ends of the true triangle shape, we will not lose much power in testing against the background hypothesis. At the same time, if x_0 is not a true binding site, then such a choice of L and R will prevent us from pooling signals that may be caused by nearby binding sites, so that we will not declare too many false positives.

If ϵ_i is stationary but not uncorrelated, with marginal variance σ^2, then under assumptions (1) and (2), $E\|\epsilon - \hat{\epsilon}\|^2 = E\|\epsilon\|^2 - E\|\hat{\epsilon}\|^2 = (L + R + 1 - \mathrm{tr}(H\Sigma))\sigma^2$, where $\Sigma = E(\epsilon\epsilon')/\sigma^2$ is the correlation matrix of ϵ. $E(\hat{\sigma}^2) = \sigma^2(L + R + 1 - \mathrm{tr}(H\Sigma))/(L + R + 1 - d)$, which depends on L and R, and which is not an unbiased estimate of the marginal variance σ^2. In this situation, we continue to choose L and R with minimum $\hat{\sigma}^2$. A simulation study in Section 5.3 suggests that this choice still produces sensible results.

Sometimes, ChIP-chip may produce an enriched region as a plateau of high values instead of a peak. In this case, our method can still detect such a region, because the truncated triangle shape model can fit such plateau shapes with very flat slopes. Occasionally, some probes may fail to function normally during the ChIP-chip experiment. Such dysfunctional probes may produce overly small or large signals. The truncated triangle shape model enables us to detect and remove such probes as outliers.

4.2 *Peak Recognition Algorithm*

(i) 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 prespecified and the default value is 200).

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

(iii) 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 prespecified, and the default numbers are 300 bp and 1500 bp, respectively). Then choose the (L, R) that gives 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 variance of x.

(iv) 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 variance as described in Step iii. Then, among the current local maximum probe and its neighbors, choose the probe with the smallest residual variance to identify the best-fitted triangle shape. We mark this probe as a potential binding site.

(v) For any local maximum probe other than the above marked probe within the range of this best-fitted triangle shape, we compare the fitted value of the best-fitted triangle and the fitted value of the triangle centered at this local maximum probe. If the difference between the two fitted values at this local maximum probe is less than a threshold (which is a factor times the standard deviation of the residuals of the best-fitted triangle, and the default factor is 1.5), then this local maximum probe is said to be explained by the best-fitted triangle and it is marked as nonpeak.

(vi) Among all the local maximum probes still not marked, choose the local maximum probe with the largest signal. Then go back to Step iii. Stop the algorithm if all the local maxima are marked.

4.3 *Peak Testing*

For a potential binding site x, suppose the truncated triangle shape fitted at x covers n probes. Let $Y_1, Y_2, \ldots, Y_n$ 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 by pooling these n probes. We use the following test statistic: $\bar{Y}_n = \sum_{i=1}^{n} Y_i/\sqrt{n}$. A similar method is proposed by Buck, Nobel, and Lieb (2005).

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

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

where γ^2 is the marginal variance $\mathrm{Var}(Y_i)$, and f is the autocorrelation factor. Both can be estimated from the data. Specifically, we can first calculate the marginal standard deviation of the whole sequence of signals. Then we remove those signals that are above a threshold (default value is 2.5 times the marginal standard deviation). After that we estimate γ^2 and f based on the remaining signals. Because the true peak shapes only occupy small portions of the whole sequence, and the vast majority of the signals are background noises, such a procedure gives reasonable estimates of γ^2 and f.

We calculate 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 peak shapes by thresholding the p-value (the default threshold is 1%).

5. Software, Results, and Related Issues

5.1 *Software*

A software named `Mpeak` has been developed for model-based peak recognition (as well as multiresolution peak

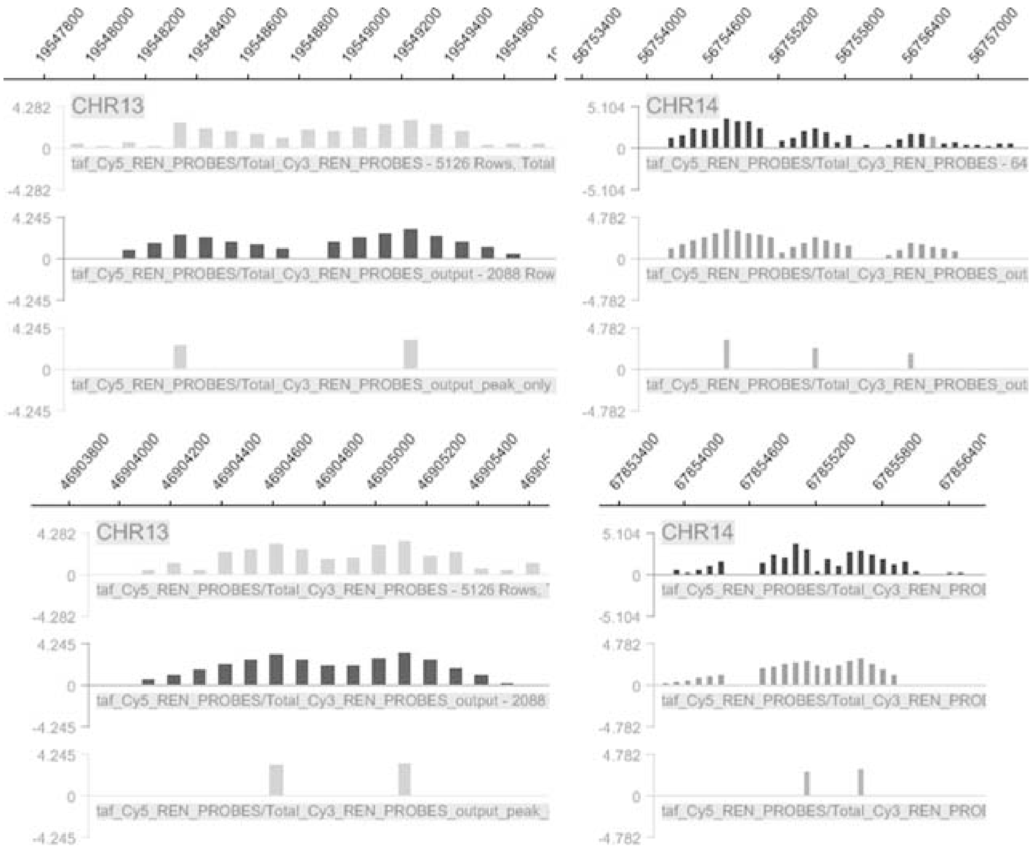

Figure 2. Top row: original data. Middle row: fitted data. Bottom row: peak positions.

tree representation to be described in Section 6). The software and the source code are free to download from `www.stat.ucla.edu/~zmdl/mpeak`. The algorithm takes less than 1 minute to analyze a genome long sequence on a regular PC.

5.2 *Results on Real Data*

Kim et al. (2005) conducted a ChIP-chip experiment for identifying the promoter regions in the entire human genome. They used probes of 50 bp to tile the nonrepetitive sequence of the whole human genome. The spacing between two adjacent probes was 100 bp. Antibodies targeting four different proteins, i.e., the TAF1 subunit of the transcription factor IID (TFIID), RNA polymerase II (RNAP), localized acetylated histone H3 (AcH3), and methylated histone H3 lysine residue 4 (MeH3K4), as used to immunoprecipitate DNA segments bound by the proteins in the human primary fibroblast IMR90 cells. Meanwhile, a control pool of DNA segments was added and dyed with a different color. The log ratio of the signals for the two fluorescent dyes was extracted, displayed by the `SignalMap` software of NimbleGen company (Madison, Wisconsin), and analyzed by our algorithm. The reader is referred to Kim et al. (2005) for biological discoveries and validations.

Figure 2 shows 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. Among all the detected peaks in this data set, the mean of the R^2 statistics is 0.82, with a standard deviation 0.22 (model-based outlier removal is performed before computing R^2).

As to the ranges covered by the fitted triangle shapes, the mean is 918 bp, and the standard deviation is 416 bp. The minimum allowable value of L and R is set at the default value 300 bp.

5.3 *Simulation Study: Autocorrelation and Minimum Range*

To examine the issues of autocorrelation and the minimum allowable value of L and R, which determines the resolution of the algorithm, we conduct a simulation study. We generate a long sequence of signals, with 120 enriched regions, separated by background signals. In each enriched region, there are two peak shapes that are close to each other. The distance between the two peak probes (i.e., the two binding sites) in each region is set at 700 bp. The left and right ranges of the peak shapes are both 300 bp. The true signal values of the two peak probes are either 2 or 2.5. The shape of a peak can be either triangular or double exponential. For triangular shape, the true signal values of the probes fall linearly from the value of the peak probe to 0 at the two ends of the range. For double exponential shape, the true signal values of the probes fall exponentially from the value of the peak probe to 0.01 at the two ends of the range. The distance between consecutive probes can be 30, 50, and 100 bp. Therefore, there are 2 peak values × 2 shapes × 3 spacings = 12 types of regions. For

Table 1
Results of Mpeak *on simulated data*

ρ	Minimum allowable range	No. of regions 1 peak detected	No. of regions 2 peaks detected	No. of regions >2 peaks detected	No. of regions no peak detected	No. of false peaks in background
0	100	6	108	6	0	64
	300	13	107	0	0	38
	500	24	93	0	3	24
0.2	100	12	103	4	1	49
	300	10	106	2	2	32
	500	23	88	0	9	17
0.5	100	19	92	6	3	34
	300	28	84	3	5	49
	500	38	66	0	16	34

There are 120 enriched regions separated by background signals. Each enriched region has two peaks. The observational errors and background data follow a first order autoregressive model. The autocorrelation ρ takes values in {0, 0.2, 0.5}.

each type of region, we simulate 10 replicates. So there are a total of 120 regions, with 240 peaks.

The additive observational errors and background signals are assumed to follow a stationary Gaussian autoregressive process, $\epsilon_i = \rho\epsilon_{i-1} + \sqrt{1-\rho^2}\delta_i$, where $\delta_i \sim N(0,\ 0.5^2)$ independently. The marginal standard deviation of this autoregressive process is 0.5. Between every two consecutive enriched regions, there are 1000 probes whose signals follow the background noise model.

Such twin peaks shapes can arise in the situation where two modified histone-binding sites exist in proximity around a promotor. Such shapes can be interesting to biologists and it is important to resolve the two peaks.

Table 1 shows the results of Mpeak under different autocorrelations with different minimum allowable values for R and L. The threshold for p-value is set at a default value 1%. When the minimum value of R and L is 100, there are slightly more false positives, and slightly fewer false negatives. When the minimum value of L and R is 500, the minimum total range $R + L$ is $500 \times 2 = 1000$, which is greater than the distance between the two peak probes, which is 700. In this case, Mpeak still shows reasonable performance. As to the autocorrelation, even when it is as high as 0.5, Mpeak still performs reasonably. Results in Table 1 are to be compared with results in Table 2 in the next section.

6. Kernel Smoothing and Multiresolution Peaks

As a nonparametric alternative to the model-based peak detection method, one can convolve the probe signals with a smoothing kernel function, such as uniform or Gaussian density function. Then one can identify the local maxima of the smoothed signals, and test the significance of these local maxima against a background model. Such methods have been proposed by Glynn et al. (2004) and Buck et al. (2005). The ChIPOTle software of Buck et al. (2005) uses a uniform kernel function and assumes Gaussian white noise for background signals.

Table 2 shows the results of kernel smoothing using ChIPOTle on the same simulated data as described in Section 5.3, where the half-window size takes values in {100, 300, 500}. The threshold for p-value is set at default value 1%, the same as Mpeak.

Table 2
Results of kernel smoothing on simulated data

ρ	Half window size	No of regions 1 peak detected	No of regions 2 peaks detected	No of regions >2 peaks detected	No of regions no peak detected	No of false peaks in background
0	100	0	117	3	0	600
	300	0	55	65	0	409
	500	57	56	7	0	433
0.2	100	1	114	5	0	1127
	300	5	67	48	0	913
	500	54	53	13	0	919
0.5	100	0	119	1	0	1974
	300	5	66	49	0	1874
	500	68	41	11	0	1892

There are 120 enriched regions separated by background signals. Each enriched region has two peaks. The observational errors and background data follow a first-order autoregressive model. The autocorrelation ρ takes values in {0, 0.2, 0.5}.

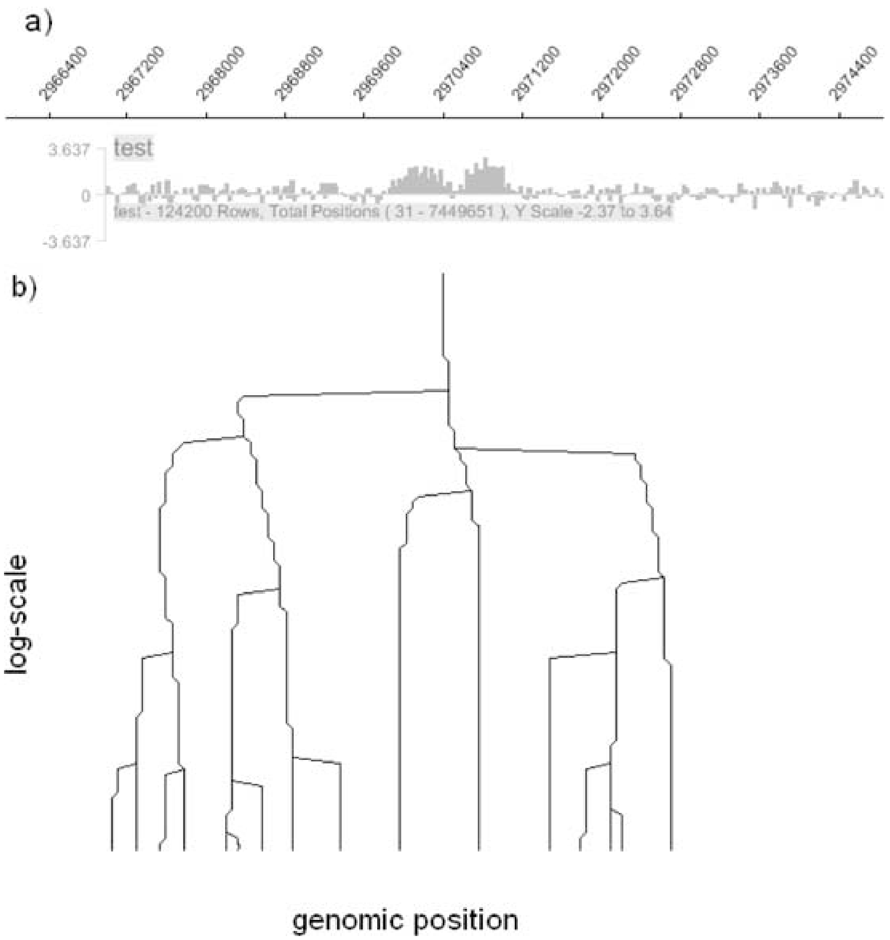

Figure 3. Multiresolution peak-tree. (a) Signals. (b) The trajectories of the local maxima over scales.

The smoothing method appears to be sensitive to the choice of bandwidth or window size. When the half window size is 100, it performs well, although there are more false positives in the background. If we increase the bandwidth to 300, the method often identifies more than two peaks in an enriched region. At a half window size 500, the method often identifies only one peak in an enriched region, and the identified peak actually corresponds to the valley, because kernel smoothing does not recognize the local shape. This is also the reason that it declares more false positives in the background than `Mpeak`, which fits fewer triangle shapes than the number of local maxima identified by smoothing. Also, the smoothing method such as `ChIPOTle` assumes white noise background, so that it can declare more false positives in the background when the autocorrelation is high.

To further illustrate the issue of bandwidth, we borrow the insight from the scale space theory (Witkin, 1984) in computer vision. We convolve the original signal $Y(x)$ with Gaussian kernel $G_s(x)$ for the whole range of standard deviations or scales $s \in [s_{\min}, s_{\max}]$ (the default range is [50, 700] in our implementation). For each s, we identify the local maxima of $Y(x) * G_s(x)$. If we plot each local maximum as a point in the joint space of (x, s), then we get the trajectories of these local maxima across scales. See Figure 3 for an illustration. Clearly, a local maximum exists within a range of scales, and two neighboring local maxima can merge into one local maximum if we keep increasing the scale s. This leads to a tree structure for organizing the multiresolution local maxima. This further illustrates the need for adaptive bandwidth selection. In particular, for two neighboring local maxima that are to be merged into a single maximum at scale s, we need to decide whether the local data should be described by two local maxima at scales below s, or be described by one single maximum at scales above s. We will investigate this issue in future work. We believe this will lead to a useful alternative to our model-based method.

`Mpeak` performs adaptive scale selection by fitting the truncated triangle shapes for all the allowable combinations of R and L.

While there is a bandwidth selection problem with the nonparametric smoothing method, if the peaks are well separated relative to the bandwidth, the smoothing method generally works well.

7. Replicates

The ChIP-chip experiments can be replicated to produce multiple sequences of signals. To analyze such replicated data, one simple method is to take the average of the replicates,

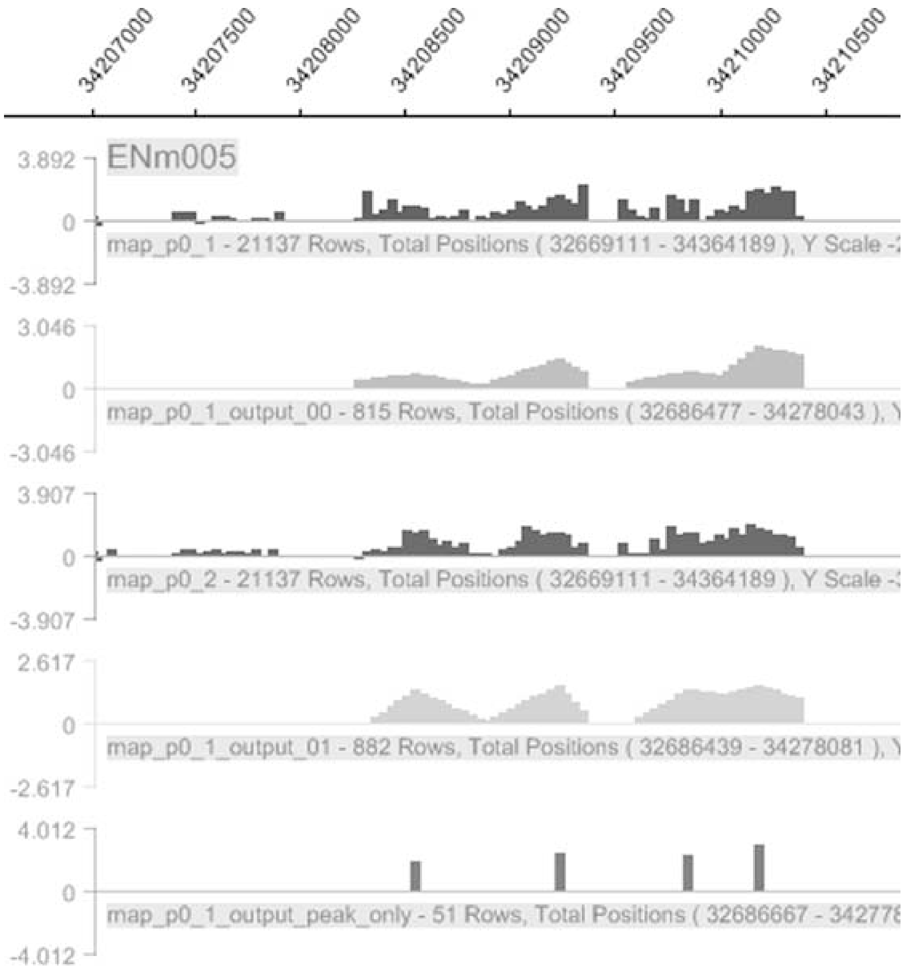

Figure 4. Replicates. The first and third rows are observed data for two replicates. The second and fourth rows are fitted shapes. The bottom row displays the positions of peaks.

and run Mpeak on the averaged signals. Another method is to run Mpeak on each replicate, and then merge the results. A more principled method is as follows. Around each probe position, and for each pair of (L, R), we fit a separate triangle shape model for each replicate, where each fitted triangle has its own intercept and slopes. Then we average the residual variances obtained from all the replicates. After that, we use the averaged residual variances to identify the positions and ranges of the potential binding sites, following the same scheme as described in Section 4.2.

Figure 4 illustrates the method using real data. The first and third rows are observed data for two replicates. The second and fourth rows are fitted shapes. The bottom row displays the positions of peaks.

We would also like to refer the reader to Li, Meyer, and Liu (2005) and Ji and Wong (2005) for analyzing replicate data. Both methods require replicates to estimate the variance of the signal intensity of each probe position across different experiment conditions. Li, Meyer, and Liu (2005) estimates the probability of a probe belonging to an enriched region using a hidden Markov model and averages the probability over the replicates. Li, Meyer, and Liu (2005) uses a t-test-like probe-level statistic to identify probes that are statistically different in different experiment conditions. Unlike our method, these two methods identify enriched regions instead of pinpointing the peaks in the signals.

8. Discussion

In the future work, we need to extend the model by relaxing assumptions such as the Poisson distribution of the cut points and the additive errors in the probe signals. We should also further develop both model-based method and nonparametric methods. In particular, in the model-based method, the model should be able to account for more complex shapes. In the nonparametric method, we should develop an automatic bandwidth selection method for peak finding.

Acknowledgements

The authors thank the two referees, the associate editor, and the editor for their insightful comments that have greatly improved both the content and the presentation of this article.

References

Buck, M., Nobel, A., and Lieb, J. (2005). Chipotle: A user-friendly tool for the analysis of chip-chip data. *Genome Biology* **6:R97**.

Dudoit, S., Yang, Y., Callow, M., and Speed, T. (2000). Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. Technical report, Department of Biochemistry, Stanford University School of Medicine, California.

Glynn, E., Megee, P., Yu, H., Mistrot, C., Unal, E., Koshland, D., DeRisi, J., and Gerton, J. (2004). Genome-wide mapping of the cohesin complex in the yeast saccharomyces cerevisiae. *PLoS Biology* **2:E259,** 1325–1339.

Ji, H. and Wong, W. (2005). Tilemap: Create chromosomal map of tiling array hybridizations. *Bioinformatics* **21,** 3629–3636.

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

Li, W., Meyer, C., and Liu, X. (2005). A hidden markov model for analyzing chip-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics* **21,** i274–i282.

Lockhart, D., Dong, H., Byrne, M., Follettie, M., Gallo, M., Chee, M., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology* **14,** 1675–1680.

Ren, B., Robert, F. Wyrick J. J., et al. (2000). Genome-wide location and function of DNA binding proteins. *Science* **290,** 2306–2309.

Witkin, A. (1984). Scale space filtering: A new approach to multi-scale description. In *Image Understanding*, S. Ullman and W. Richards (eds), 79–95. Norwood, New Jersey: Ablex.

Received October 2005. *Revised November* 2006.
Accepted November 2006.