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 $\alpha$. 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 $Y(x) = \log(C_{\text{Cy5}}/C_{\text{Cy3}})$. It measures the relative abundance of good fragments that contain $x$. 
\[ p_R(x) \]
\[ = \Pr(\text{no cut between } x \text{ and the nearest good binding site to the right}) \]
\[ = \sum_{i=m+1}^{M+1} \left( \prod_{j=i-1}^{m+1} (1 - q_j) \right) q_i \exp\left\{ - \int_x^{B_i} \lambda(s) ds \right\}. \]  

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 \( Y(x) = \log(Cy5/Cy3) \), 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:

\[
Y(x) = \log(Cy5/Cy3) = c - b[B - x]^+ - a[x - B]^+ + \epsilon(x), \tag{10}
\]
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

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

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

$$(\hat{c}, \hat{b}, \hat{a})' = (X'X)^{-1}X'Y,$$

$$\hat{\sigma}^2 = \|Y'Y - Y'X(X'X)^{-1}X'Y\|^2 / (R + L + 1).$$

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 $x$. 

10
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 appears 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 $R^2$


