Model-based deconvolution of genome-wide DNA binding

David J. Reiss*, Marc T. Facciotti and Nitin S. Baliga

Institute for Systems Biology, 1441 N. 34th St. Seattle, WA 98103-8904, USA

Received on August 16, 2007; revised and accepted on November 27, 2007
Advance Access publication December 1, 2007
Associate Editor: Martin Bishop

ABSTRACT

Motivation: Chromatin immunoprecipitation followed by hybridization to a genomic tiling microarray (ChIP-chip) is a routinely used protocol for localizing the genomic targets of DNA-binding proteins. The resolution to which binding sites in this assay can be identified is commonly considered to be limited by two factors: (1) the resolution at which the genomic targets are tiled in the microarray and (2) the large and variable lengths of the immunoprecipitated DNA fragments.

Results: We have developed a generative model of binding sites in ChIP-chip data and an approach, MeDiChI, for efficiently and robustly learning that model from diverse data sets. We have evaluated MeDiChI’s performance using simulated data, as well as on several diverse ChIP-chip data sets collected on widely different tiling array platforms for two different organisms (Saccharomyces cerevisiae and Halobacterium salinarium NRC-1). We find that MeDiChI accurately predicts binding locations to a resolution greater than that of the probe spacing, even for overlapping peaks, and can increase the effective resolution of tiling array data by a factor of 5 or better. Moreover, the method’s performance on simulated data provides insights into effectively optimizing the experimental design for increased binding site localization accuracy and efficacy.

Availability: MeDiChI is available as an open-source R package, including all data, from http://baliga.systemsbiology.net/medichi.

Contact: dreiss@systemsbiology.org

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Localization of the genomic targets of DNA-binding proteins (e.g. transcription factors (TFs), histones, etc.) is central to understanding gene regulation and chromatin organization. Chromatin ImmunoPrecipitation (ChIP) is an experimental protocol that can be used to enrich protein–DNA complexes using antibodies against a protein of interest. The chromatin in these enriched complexes are then randomly fragmented and hybridized against a tiling microarray (chip). This combined procedure, hybridizing a ChIP sample against a DNA microarray, is often referred to as ChIP-chip or ChIP-to-chip (Ren et al., 2000; Lee et al., 2002).

The enriched DNA fragments are of variable length, usually ~300–500 bp long—often far longer than the spacing (tiling array resolution; hereafter TAR) of the microarray probes. Thus, each binding event influences the intensities of multiple probes adjacent to the one containing the binding site (TFBS). Most methods for detecting binding sites in such data identify contiguous windows of significantly enriched probes (e.g. Buck et al., 2005; Ji and Wong, 2005; Johnson et al., 2006). Therefore, regardless of the TAR or the accuracy of these detection methods, the locations of the identified TFBS are still limited to windows of at least the average length of the enriched DNA fragments.

Of course, the accuracy of all TFBS prediction methods using ChIP-chip data (including the one we will describe here) depends upon proper normalization of the data (e.g. normalization versus a reference sample of sonicated chromatin, or via models of probe-specific hybridization, such as those of MAT (Johnson et al., 2006) or others (Cambon et al., 2007). Some methods for detecting TFBS integrate normalization into their procedure; others assume that the normalization is performed ahead-of-time. However, no method is immune to poor data quality or improper normalization (see e.g. Euskirchen et al., 2007 for a review).

Recently, Qi et al. (2006) developed the Joint Binding Deconvolution (JBD), a method that learns a Bayesian graphical model of joint binding events in normalized ChIP-chip data to infer posterior-binding probabilities (and strengths) at resolutions (~30 bp) better than that allowed by previous methods—and better than the TAR itself. Uniquely, the method uses a model for peak profiles derived from experimentally-measured DNA fragment lengths. JBD requires multiple replicates or previous experiments (in order to parameterize the noise model), as well as DNA fragment lengths, measured separately for each sample via electrophoretic analysis. Additionally, as it is a Bayesian inference procedure, JBD requires several user-defined prior parameters, e.g. two hyper-prior parameters for the likelihood of binding (which will vary depending on the type of DNA-binding molecule). JBD’s generated binding distributions may be used as a prior constraint for motif detection (Qi et al., 2006). However, their interpretation as discrete TFBS necessitates post-processing (via a user-defined threshold), which can decrease the effective resolution and information content of the predictions.

In this article, we introduce MeDiChI, a regression-based method that learns a generative model of joint (multiple, potentially overlapping) binding events in normalized ChIP-chip data. It estimates the positions of likely binding sites via constrained, penalized least squares regression. In Section 2,
we derive a model for the ChIP-chip profile of an individual binding event, which becomes the deconvolution kernel in a joint model of multiple binding events. We show how this model is learned from data, and demonstrate (Section 3) that our procedure is robust, and is able to correctly predict locations of binding events at high resolution from data of widely varying quality and resolution: original 

*H. salinarium*

NRC-1 general transcription factor (GTF) TFBd binding data; previously-published data for the yeast TF Gcn4 (Pokholok et al., 2005); and simulated binding data. We also perform an in-depth comparison of MeDiChI with other popular ChIP-chip analysis methods, including JBD (Qi et al., 2006). Finally, we show that the application of MeDiChI to simulated data reveals important considerations for the design of future ChIP-chip experiments.

## 2 METHODS

### 2.1 The model of a single binding event

The ChIP-chip assay has been described previously in detail (Ren et al., 2000; Lee et al., 2002), and we depict the protocol and primary aspects of our single binding event model (on a tiling array) in Figure 1. We introduce the following physical parameters (all genomic units are in base-pairs):

1. the probe length, $2 \times l_P$ (assumed to be the same for all probes),
2. the footprint $2 \times \lambda$ of the DNA binding protein on the DNA,
3. the distribution $f(l)$ of enriched, sonicated DNA fragments of length $l$ cross-linked to the TF,
4. the intensity $I(l)$ of a labeled DNA fragment of length $l$ and
5. the conditional probability $\Phi(v)$ of hybridization of each fragment to those probes to which it has contiguous complementarity by $\psi$ bp.

An enriched DNA fragment of length $l$ containing a binding site at genomic coordinate $x_0$ must begin at a coordinate somewhere on the interval $[x_0 + \lambda - l; x_0 - \lambda]$. Such a fragment can hybridize with any probe that intersects the interval $[x_0 + \lambda - l; x_0 - \lambda + l]$. These probes’ center coordinates will then lie in the range $[x_0 + \lambda - l - l_P; x_0 - \lambda + l + l_P]$. Therefore, the fragment has the following conditional probability of binding to a probe centered at position $x$:

$$P_a(x - x_0) = \frac{1}{l - 2\lambda} \sum_{j=1}^{l/\lambda} \Phi \left( \sum_{j=1}^{l/\lambda} H(j, x - x_0) \right).$$

$$H(j, x - x_0) = \begin{cases} 1 & \text{if } j \in [x - x_0 - l_P; x - x_0 + l_P] \\ 0 & \text{otherwise} \end{cases}$$

The final intensity profile is obtained by integrating the fragment length distribution $f(l)$ and intensities $I(l)$ over all fragment lengths $l$, weighted by the probability $P_a(x - x_0)$ (Equation 1) of hybridizing with the probe at $x - x_0$:

$$K(x - x_0) = \sum_{l} f(l) I(l) P_a(x - x_0).$$

$K(x - x_0)$ is a symmetric, inverted, nearly-triangular profile, with broader wings contributed by longer DNA fragments. This profile is used as a kernel for subsequent deconvolution via linear regression.

**Model parameters:** The parameter with the greatest influence on the shape of $K(x - x_0)$ is the distribution of fragment lengths $f(l)$, which we parameterize using a Gamma distribution, $\Gamma(\alpha, \beta)$. We use a Sigmoidal distribution, $1/(1 + \exp[-(v - v_0) \times v_0])$ for $\Phi(v)$. We find that the effects of $f(l)$ and $\Phi(v)$ on the profile shape are highly covariant with that of $I(l)$, and therefore these first two are fixed at canonical values: $\alpha = 0.1$; $v_0 = 15$; $v_0 = 0.1$ (making $\Phi(v)$ a step function $[1$ if $v \geq 15$; 0 otherwise]). Any difference between these settings and the ‘real’ values can be captured by changes in the scale $\alpha$ and shape $\beta$ of $f(l)$. Practically, while $\lambda$ can strongly influence the profile’s shape (by dulling the point of the peak profile), it can also be captured by imposing a lower-limit on $f(l)$. Thus only $f(l)$ is necessary to uniquely define $K(x - x_0)$ for a given data set (for which $l_P$ is given). $f(l)$ for each ChIP-chip sample may be obtained via electrophoretic analysis, such as via a DNA BioAnalyzer (Agilent Technologies, Santa Clara, CA) (Qi et al., 2006). Alternatively, $f(l)$ (and, optionally $l$) may be learned directly from the tiling array data itself, as described in Section 2.6.

### 2.2 A constrained linear model of multiple binding events

Equation (2) gives the intensities $K(x - x_0)$ of probes at distance $x$ from the location $x_0$ of an individual binding event. We now discuss the joint model of all functionally independent binding events for a given TF simultaneously across the genome. The measured intensity $y_i$ of probe
i centered at chromosomal location $x_i$ is the sum of intensities arising from each of the binding events at nearby locations $x_j$:
\[ y_i = \beta_0 + \sum_{j=1}^{N_b} \beta_j \text{K}(x_i - x_j) + \varepsilon_i \]  

(3)

$\varepsilon_i$ is the error term for probe $i$, which includes random noise and systematic (possibly correlated) variance which is not explained by our model (such as cross-hybridization; see below). The parameters $\beta_j$ quantify the intensity of the binding event at $x_j$, which we note is an (unknown) function of various factors, including the fraction of cells with the binding event occurring at the time of cross-linking, the binding affinity at that location, the efficiency of cross-linking or of IP of the protein–DNA complex, and the efficiency of IP, amplification and hybridization of the DNA fragments containing the site. However, for a given TF we can expect that each $\beta_j$ is related to the TF’s affinity for location at $x_j$.

$\mathbf{K}_b = \mathbf{K}(x_i - x_j)$ is a $N \times N_b$ matrix containing the set of predictors for the $N_b$ potential binding locations. $N_b$ is defined by the desired resolution $\tau_b$. We have found (Section 3) that a reasonable default for this quantity is on the order of $\tau_b \approx \max(\lambda \times \gamma/10, 10)$.

The coefficients $\beta$ are solved by minimizing the residual sum of squares,
\[ \text{RSS}(\beta) \equiv \sum_{i=1}^{N} y_i^2 - \sum_{j=1}^{N_b} (y_i - \beta_0 - \sum_{j=1}^{N_b} \beta_j \text{K}(x_i - x_j))^2 \]  

(4)

Here we have assumed independence in the observations $y$ and identical and uncorrelated errors $\varepsilon$. We are thereby ignoring secondary effects such as (1) nonlinearity (saturation) in the microarray intensities, which could potentially be alleviated through the use of a generalized linear model (Bonneau et al., 2006); (2) independence of the relevant functional DNA-binding unit (where a dimer, for example, is considered one independent functional-binding unit); (3) cross-hybridization and probe sequence-related systematics, which is usually (mostly) removed through normalization, or via a model such as that proposed by Johnson et al. (2006); and (4) non-constant variances $\sigma$, which can be accounted for with weighted least squares.

Finally, binding events generate only positive signal, so we enforce, for each coefficient $\beta_i$,
\[ \beta_i \geq 0 \]  

(5)

### 2.3 Constrained variable selection via positive Lasso

In most cases, for the regression defined above, we will have $N \ll N_b$. In addition, many of the predictors $K_j$ are highly-correlated and therefore redundant. Therefore, to avoid overfitting, we wish to select a sparse model defined by a small subset $N_b \ll N_b$ of the complete set of variables. There are many methods available for variable subset selection or coefficient shrinkage, and of these, we chose the Lasso (Hastie et al., 2001), which applies an $L_1$ penalty to the coefficients. The $L_1$ constraint results in a sparser solution, which is what is desired here, than other commonly-used penalties, such as $L_2$ (ridge). Under the Lasso constraint, we update Equation (4) to:
\[ \hat{\beta}_{\text{Lasso}} = \arg \min_{\beta} \sum_{i=1}^{N} (y_i - \sum_{j=1}^{N_b} \beta_j \text{K}(x_i - x_j))^2, \]  

subject to $\sum_{j=1}^{N_b} \beta_j \leq t$ and $\beta_j \geq 0$.

(6)

We have also added the positivity constraint of Equation (5). We choose a value for the ‘tuning parameter’ $s \equiv t/\sum_{j=1}^{N_b} \beta_j$ in Equation 6 to select a model $M_s$ that increases in complexity (and bias) as $s$ increases from zero (the null model $M_0$) to one [the OLS estimate $M_{\text{OLS}}$ Equation (4)]. Least angle regression (Lars; Efron et al., 2004), an efficient procedure for generating piecewise-linear solutions to Equation 6 over small intervals of $s \in [0; 1]$; produces a set of $h_{\text{Lasso}}$ models with increasing complexity as $s$ increases. In practice, we limit the number of Lars steps computed $k_{\text{max}} \approx N_b / 4$, corresponding to the model of greatest complexity that is considered. $M_{\text{max}}$. In addition, we modified the authors’ implementation (Efron et al., 2004) to enforce the additional positivity constraint in the coefficients (Equation 6).

The Lasso constraint has been effectively utilized for other bioinformatics problems such as learning regulatory influences from microarray data (Bonneau et al., 2006; van Someren et al., 2006) and measuring genomic copy number alterations in tiling array data (Huang et al., 2005). The Lasso has also been previously applied to the deconvolution of mass spectra (Du and Angeletti, 2006), and 2D images (Ting et al., 2006).

### 2.4 Model selection via BIC

Assuming that the underlying model is correct, parametric model selection criteria such as Akaike information criterion (AIC), or Bayesian information criterion (BIC) can be more accurate and less costly than cross-validation (Zou et al., 2004). When $\log (N) \gg 2$, BIC tends to penalize complex models more heavily than AIC (Hastie et al., 2001), which is precisely what is desired here. In addition, it has been shown that BIC is typically more consistent in selecting the true model if it is among the list of candidate models (Zou et al., 2004). These motivations led us to choose BIC for model selection:
\[ \text{BIC}(s) = \text{RSS}(\hat{\beta}_s)/s^2 + \log(N) \Delta(s) \]  

(7)

(Zou et al., 2004), for a given set of parameters $\hat{\beta}_s$ selected for model $M_s$ at Lasso shrinkage parameter $s$. Here, $\Delta(s)$ can be estimated by the mean squared error of a high-complexity model (Hastie et al., 2001); we use $\text{RSS}(\hat{\beta}_{\text{max}}/N)$; and $\Delta(s)$ is the effective number degrees of freedom of model $M_s$. It was empirically found (Efron et al., 2004; Zou et al., 2004) that $\Delta(s)$ is well-approximated by the step number $k_{\text{Lasso}}$ of the most complex model in the Lars sequence that has the same number of non-zero parameters as does $M_s$. Model selection via BIC involves simply identifying the model $M_s$ that minimizes BIC$(s)$.

The final selected model $M_s$ consists of a limited set of nonzero coefficients $\beta_s$, defining the fitted intensities of a set of binding sites at their predicted high-resolution locations. However, we find (via simulations; see below) that Lasso often shrinks the nonzero coefficients $\beta_s$ to slightly below their optimal values, thereby underestimating their intensities. We therefore update these by estimating the final solution to Equation (4) (for only the nonzero coefficients) under the positivity constraint [Equation (5)] via quadratic programming (Goldfarb and Idnani, 1983).

### 2.5 Deconvolution of an entire ChIP-chip data set

Because the size of the kernel matrix $K$ increases with the square of the size of the genomic region being modeled, we break up the deconvolution of an entire ChIP-chip data set into smaller (e.g. 30000 bp) overlapping segments, so that they can be deconvoluted at high resolution (e.g. 10 bp) with limited memory requirements. These resulting fits are then combined into a single set of peak locations and intensities for each chromosome. Such a procedure is easily distributed to multiple processor cores or on a computing cluster.

Additionally, we note that this procedure is performed simultaneously on multiple technical and/or biological replicates without signal averaging.

### 2.6 Learning peak profile parameters

The accurate model of a single, isolated peak [Equation (2)] for each data set is important to enable precise predictions of correct binding sites. The model of Equation (2) depends upon several parameters, which can be...
heteroskedastic and spatially-correlated characteristics of the original data set. The iterative procedure begins with reasonable starting values for the parameters, and repeats the following:

1. Compute a peak profile \( K(x - x_0) \) [Equation (2)] using the current parameter values.
2. Apply the Lars/BIC procedure described above to identify the \( N_p + N_A \) brightest peaks matching the profile.
3. Compute, for each of the \( N_p + N_A \) peaks, the mean RSS for the probes within \( \sim 1000 \) bp of its center, and identify the \( N_p \) of these peaks with the lowest mean RSS (presumably isolated peaks).
4. Compute the total mean RSS for all probes surrounding the \( N_p \) isolated peaks, and update the model parameters using the Simplex method (Nelder and Mead, 1965) to minimize this score.

This procedure usually requires \( \sim 30 \) iterations.

2.7 Assessment of model uncertainty

Due to the complexity of the Lars procedure (and selection via BIC), we cannot compute parametric estimates of the uncertainties in the inferred binding site parameters. Therefore, we use the bootstrap to estimate the uncertainties in either (1) which coefficients are selected by the Lars/BIC (and thus in the coordinates of the predicted binding sites), or (2) their significance. Further, the effects of the model’s i.i.d. assumptions (describe above) on the predictions can be tested.

For this article, we chose the wild bootstrap (Liu, 1988) data generation procedure for this analysis. The advantage of this data generation procedure is that its simulated data sets mimic the heteroskedastic and spatially-correlated characteristics of the original data. For bootstrap \( j \in [1, N_{\text{boot}}] \), new residuals \( \tilde{\gamma} \) are sampled from an asymmetric two-point distribution (Mammen, 1993),

\[
\tilde{\gamma}_j = \tilde{\gamma}_i \times \begin{cases} 
(\sqrt{5} + 1)/2 & \text{with probability } (\sqrt{5} - 1)/(2\sqrt{5}) \\
-(\sqrt{5} - 1)/2 & \text{with probability } (\sqrt{5} + 1)/(2\sqrt{5})
\end{cases}
\]

where the data residuals \( \tilde{\gamma} \) [Equation (4)] are computed from model \( \mathcal{M}_i \) (parameterized by \( \tilde{\beta} \)). A sample data set \( \tilde{\gamma}_j \) is generated, for which a set of \( \tilde{\beta}_j \) coefficients are selected using the Lars/BIC procedure described above. The sampled data can mimic either (1) the intensities of the original data \( \tilde{\gamma}_j = \beta_j + \tilde{\gamma}_j \) (or 2) the residuals alone \( \tilde{\gamma}_j = \tilde{\gamma}_j \). In the first case, a distribution of predicted binding locations may be computed via a kernel-based density estimator of the resulting \( \tilde{\beta} \). In the second case, the \( p \) value for binding site \( \tilde{\beta} \) is estimated from the CDF of all nonzero \( \tilde{\beta} \)s over all bootstraps:

\[
p_j = \frac{\sum_{i=1}^{N_{\text{boot}}} \sum_{\{\tilde{\beta}_i > \tilde{\beta}_j\}} \Delta(x)}{N_{\text{boot}} \times |\{\tilde{\beta}_i\}|},
\]

where the \( \tilde{\beta}_k \) are among the \( |\tilde{\beta}_i| \) non-zero \( \tilde{\beta} \) detected (in the entire data set) for bootstrap \( j \), and \( \Delta(x) \) equals 1 when \( x \) is true; 0 when \( x \) is false. The computational time required for the \( N_{\text{boot}} \) bootstraps is significantly less than \( N_{\text{boot}} \times \) (the time for a single non-bootstrap MeDiChI estimate) because many computationally intensive operations (e.g. construction of the kernel matrix \( K \)) need only be performed once.

3 RESULTS

We assessed the performance of MeDiChI deconvolution using three real-world data sets: ChIP-chip data for the \( H. \ salinarium NRC-1 \) TFBd from two microarray platforms of varying resolution (Facciotti et al., 2007) (Section 3.1); and previously-published \( S. \ cerevisiae \) Gcn4 binding data (Pokholok et al., 2005) (section 3.2). We also tested MeDiChI on simulated data (section 3.3), to evaluate the accuracy and limitations of the algorithm on data of varying resolution and quality.

3.1 MeDiChI performs well on both low-resolution and high-resolution data

Genome-wide ChIP-chip data for \( H. \ salinarium NRC-1 \) were collected for the cmyc epitope-tagged GTF TFBd, by hybridizing the same biological sample against two different whole-genome tiling array platforms. The first of these was a low-resolution (‘low-res’) \( \text{HaloSpan} \) (Facciotti et al., 2007) 500 mer PCR-product array spotted in-house in octuplicate with a TAR of 500 bp. The second array utilized a 392000-feature Nimblegen Systems (Madison, WI) array with single-stranded 50mer probes spotted in duplicate with a TAR of 13 bp (‘hi-res’). For each set, we estimated the parameters for the peak profile model (Section 2.6). The resulting profiles for both data sets are shown in Figure 2.

A sample binding site is shown in Figure 3 (a larger segment is shown in Supplementary Fig. 1). The high density of binding sites for this global regulator suggests that we may consider this data set a rather stringent test of the efficacy of the algorithm. Since the TFBSs appear, as expected, to be concentrated in the small fraction of non-coding regions (NCRs; only \( \sim 13\% \) of the entire \( H. \ salinarium NRC-1 \) genome is non-coding), we chose to use the fraction of detected binding sites in NCRs as a measure of the method’s deconvolution capabilities. (However, some TFBS will inevitably fall in coding regions, due to the high coding density of the \( H. \ salinarium NRC-1 \) genome.)

This analysis (Fig. 4) reveals that the TFBSs identified by all methods fall prevalently in NCRs, and that MeDiChI performs roughly equally well in this measure on both the high- and low-res data. We also ran \( \text{jBD} \) (Qi et al., 2006) (using default parameters and the peak profile learned directly from the data as described above), \( \text{MPeak} \) (Kim et al., 2005), and a naive threshold-based peak detection procedure (whereby TFBS are assigned to the central coordinates of probes whose IP enrichment ratio surpasses a threshold) on both data sets.
MeDiChI appears to outperform all of these methods, for both arrays, and to a greater degree as the detection threshold is decreased (thus as fainter peaks are included).

The Nimblegen array’s extremely high resolution of 13 bp allows us to consider using the peaks identified in this data set to test the TFBS localization capabilities of an algorithm on the low-res data. An all-versus-all comparison of such a comparison, for the 220 brightest peaks detected by each method (corresponding to the number of peaks detected by MPeak; the fewest of any of the methods), is shown in Table 1. We find that the binding site predictions of all four methods in the hi-res data are, on average, closer to the known gene coding regions shown below each plot. Known gene coding regions are shown below each plot. We note that we had difficulty finding parameters for JBD that resulted in (visually) satisfactory JBD predictions on all peaks in the hi-res data (e.g., the left-most peak in Supplementary Fig. 1b). Therefore, a further in-depth comparison of the two methods on data in which the JBD default parameters are known to be appropriate follows in the next section.

**Table 1.** Median distances between the 220 brightest TFBS binding site predictions from the low-res array (Rows) versus those from the hi-res array (Columns), for MeDiChI, JBD (Qi et al., 2006), MPeak (Kim et al., 2005), and basic thresholding (see text)

<table>
<thead>
<tr>
<th></th>
<th>HaloSpan (500 bp)</th>
<th>Nimblegen (13 bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MeDiChl</td>
<td>Threshold</td>
</tr>
<tr>
<td>MeDiChl</td>
<td>53.0</td>
<td>56.5</td>
</tr>
<tr>
<td>Threshold</td>
<td>87.5</td>
<td>68.5</td>
</tr>
<tr>
<td>JBD</td>
<td>62.0</td>
<td>63.0</td>
</tr>
<tr>
<td>MPeakmonospace</td>
<td>75.0</td>
<td>69.5</td>
</tr>
</tbody>
</table>

The numbers in bold indicate distances between predictions on low- and high-resolution data for a given algorithm.

3.2 MeDiChI shows superior performance on yeast Gcn4 binding data

Qi et al. (2006) performed a complete comparison of their JBD procedure against several others on a 266 bp resolution in vivo Gcn4 ChIP-chip binding array consisting of ~44000 60 mer oligonucleotide probes (Pokholok et al., 2005). They convincingly showed that their bayesian method outperformed several others (Boyer et al., 2005; Kim et al., 2005) at resolving the positions of known Gcn4 targets. We now compare the spatial resolution of TFBSs detected by MeDiChI with those of JBD (and transitively, the methods which JBD were compared against (Qi et al., 2006)) on this same data set. To ensure that the optimal JBD predictions for this data set were used, we utilized the posterior-binding probabilities and strengths (hereafter PBDs) generated by the authors http://cgs.csail.mit.edu/jbd/). JBD PBDs were discretized into binding events using the authors’ scripts with the default parameters.

TFBS predictions (examples in Fig. 6 and Supplementary Fig. 2) were compared with the 128 known Gcn4 targets (Harbison et al., 2004) after filtering all predictions to only the $n_{pred} \in \{80, 128, 200\}$ brightest across the entire genome (Table 2). A “missed” binding event was one for which none of the $n_{pred}$ brightest predictions were within 500 bp of any known location. For comparison, we also used a basic ‘thresholding’ technique in which we identified the $n_{pred}$ brightest probes in the data, and chose their center coordinates to be predicted TFBS. We also computed similar statistics for randomly-chosen probes. While both MeDiChI and JBD attained significantly better average resolutions than the thresholding method, MeDiChI appears to outperform JBD by this measure. Although the necessity of discretizing JBD’s predictions (e.g. red rectangles in Fig. 6) clearly results in loss of information contained in the PBDs (solid blue lines in Fig. 6), it reveals a problem in interpreting JBD’s PBDs as discrete binding events.
3.3 Analysis of synthetic data highlights trade-offs in experimental design

We generated synthetic ChIP-chip data, where the locations and intensities of all binding events are precisely known, to assess the accuracy of MeDiChI’s predictions on data of widely varying quality and resolution. Since the peak profile is also known, we do not use this data to test the accuracy of the profile model itself. The synthetic data were used to test the effect of probe spacing, replicate count, DNA fragment length, and binding site separation on the accuracy of MeDiChI’s predictions. The tests were divided into two classes: one to evaluate MeDiChI’s predictions on single, isolated binding events, and one to test MeDiChI’s ability to correctly separate two binding events at various separations.

We generated 7000 bp regions, each with varying tiling resolution (between 10 and 500 bp), and replicate counts (between 1 and 9). Each had zero signal except for either one or two binding events of equal intensity (peak intensities were set to one in all cases; see below), varying separation (between 50 and 700 bp), and profile width (by varying \( f(l) \) in the model of Equation (3)). Gaussian-distributed noise was then added to each probe, at a level of \( \sigma_\text{in} = \sigma_\text{out} (y_i + \sigma_\text{in}) \), where \( y_i \) is the probe intensity and \( \sigma_\text{in} \) the input noise scale (we used \( \sigma_\text{in} \in \{0.05, 0.1, 0.2, 0.3\} \)). Because the simulated peaks all had the same intensities, we considered the detection efficiency as a function of \( \sigma_\text{in} \) (effectively, signal-to-noise) rather than absolute peak intensity. Projections of the median measured resolutions of the MeDiChI predictions on these simulated data are summarized in Figure 7 and Supplementary Figure 3.

Inspection of the Figures reveals that MeDiChI increases the effective resolution of data over all tested TARs, with the greatest fractional gain for the lowest-resolution data. Not surprisingly, the resolving power of MeDiChI on pairs of binding sites (Figures 7a and b) improves for greater separations, and asymptotically approaches that for isolated sites (Fig. 7a-c).

We see that while there is a clear trade-off between TAR and number of (technical or biological) replicates, the gain seems limited when the number of replicates exceeds approximately three. The benefit of additional replicate observations is also less pronounced for higher TARs, suggesting that the parameter to consider in experiment design is actually the number of unique observations per binding site. In fact, the biggest apparent gain seems to come from increasing the

**Table 2.** Median distances and false negatives (missed events) between the \( n_\text{pred} \) brightest predicted Gen4 TFBSs, and the 128 known TFBSs (Harbison et al. 2004), for MeDiChI, JBD (Qi et al. 2006), and basic thresholding (see text)

<table>
<thead>
<tr>
<th>Method</th>
<th>( n_\text{pred} = 200 )</th>
<th>( n_\text{pred} = 128 )</th>
<th>( n_\text{pred} = 80 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>MeDiChI</td>
<td>51 ± 42; 30</td>
<td>48 ± 38; 49</td>
<td>45 ± 37; 76</td>
</tr>
<tr>
<td>JBD</td>
<td>68 ± 61; 28</td>
<td>67 ± 59; 57</td>
<td>67 ± 63; 75</td>
</tr>
<tr>
<td>Threshold</td>
<td>91 ± 74; 47</td>
<td>94 ± 95; 56</td>
<td>84 ± 81; 81</td>
</tr>
</tbody>
</table>

The value for randomly-selected tiles is 270.

**Fig. 5.** Median resolution of TFBS predictions, as a function of number of detections, on yeast Gcn4 data (Pokholok et al., 2005), for MeDiChI (using both the JBD binding site profile and ours – see text); JBD (Qi et al. (2006); using both the JBD peak filtering script and our own procedure) and threshold-based peak selection.

Median resolutions for Gen4 TFBS predictions were computed for both MeDiChI and JBD over a range of sensitivities (number of predicted TFBSs, \( n_\text{pred} \)) (Fig. 5). The MeDiChI TFBS predictions were computed using peak profiles both (1) estimated by Qi et al. (2006) for this data set, and (2) fit by our procedure directly from the data (Section 2.6). The JBD TFBS predictions were selected using both (1) the authors’ filtering script, and (2) a custom procedure that we developed. For a wide range of sensitivities, by this measure, MeDiChI appears to outperform the other methods, in particular when the peak profile learned directly from the data (Section 2.6) is used.

In general, there was a large degree of agreement between the predictions of MeDiChI and JBD, (for example, Fig. 6a and Supplementary Figure 2b), with some exceptions (e.g. Fig. 6b and Supplementary Fig. 2d). Clearly, the results such as those shown in Table 2 and Figures 6 and 5 suggest that by integrating the predictions of both methods, one could both decrease the uncertainty in binding site predictions, and reduce the number of false negatives. For example, with \( n_\text{pred} = 128 \), MeDiChI locates 18 TFBS that JBD misses, and likewise, JBD identifies 13 sites that MeDiChI misses.
TAR (and therefore the number of observations at unique locations across the binding site) than from increasing the number of observations per probe. This effect is not surprising, and is similar to what is known in imaging (e.g. observational astronomy) circles.

We also see this latter effect in a plot of median resolution (on isolated binding sites) as a function of the mean chromatin fragment length (Fig. 7a). TFBSs are localized more accurately when this length is approximately greater than the probe spacing. Thus, for lower-resolution tiling arrays, less aggressive sonication may be beneficial. This result, as above, suggests that TFBS localization accuracy increases if there are more observations at multiple locations across each observed peak. For very high-resolution arrays, this effect vanishes, although there is of course a limit to the degree of sonication that is possible. While this result is true for individual binding sites, it should be noted, however, that increasing the mean DNA fragment length will decrease the effective resolvability of two (or more) closely-overlapping peaks.

Finally, Figure 7c and Supplementary Figures 3(c and d) show the fraction of peaks that were not detected by MeDiChI (the number of false negatives) on the simulated data. On the two-peak simulations, a non-detection was defined as a case where only zero or one peak was reported. These curves explain the apparent increase in resolution for even the low-resolution data at small peak separations (Fig. 7a). The detection rate increases quickly as the separation increases, roughly independent of the TAR (except for the 500 bp, case, which is a clear outlier). For isolated peaks, the number of peaks detected (to within 1000 bp) drops from nearly 100% as the TAR decreases (Supplementary Figures 3(c and d)). The number of replicates and, even more so, the degree of signal-to-noise of the binding sites, also affects the number of false negatives.

4 DISCUSSION

We have presented a generative model the probe intensities in ChIP-chip data and an efficient, accurate, and robust method for solving that model via constrained linear regression. We demonstrated that the procedure recovers accurate binding site locations from data covering a wide range of resolutions: both simulated data covering 10–500 bp resolution, and real arrays collected for a H. salinarium NRC-1 GTF and for the yeast TF Gcn4 at very different resolutions (500, 13 and 266 bp, respectively). Our analyses suggest that the lower-resolution arrays benefit most from the resolution gain associated with the use of this algorithm, although the potential still exists for resolution gains for TARs as low as ~10 bp or even lower. We also demonstrated that the generative nature of this model can be used to produce realistic simulated ChIP-chip data, which enables the analysis of experimental design tradeoffs. For example, we showed that the utility of increasing the number of replicates beyond three is questionable, however increasing the average sonicated DNA fragment size can be beneficial for localizing binding sites using low-resolution arrays, in particular.

For localizing TFBSs in ChIP-chip data, MeDiChI clearly outperforms existing methods that cannot increase the resolution of the binding locations above that of the microarray probe spacing, (e.g. Buck et al., 2005; Kim et al., 2005). JBD (Qi et al., 2006), which resembles MeDiChI in both its model and in its ultimate goal, performs the most similarly to MeDiChI. Indeed, one may consider MeDiChI to be the frequentist alternative to the Bayesian JBD, whose predicted posterior-binding distributions are related to MeDiChI’s predictions of discrete binding locations and intensities (which may be broadened to posterior-probabilities via the bootstrap; see Section 2.7). In fact, as we presented in Section 3.2), in some instances these two algorithms complement one another and we suggest that their integration may improve overall performance in terms of accuracy, sensitivity, and specificity.

While the goals and overall model of MeDiChI and JBD are quite similar, we believe that MeDiChI exhibits several advantages that make it more practical for general use. First, MeDiChI requires only one user-defined parameter: the desired deconvolution resolution (and even for that, a reasonable default may be chosen). Second, MeDiChI includes a procedure by which the peak profile model (required by both MeDiChI and JBD) can be learned directly from the data. This enables, for example, the reanalysis previously published data where the DNA fragment length distribution f(l) was not measured. Finally, the computational cost of a non-bootstrapped MeDiChI deconvolution (once the peak profile is known) is only slightly higher than the simpler window-based algorithms listed above, and the procedure is easily distributed on multiple processor cores or a computing cluster. All of these features make MeDiChI suitable
for integration into an automated ChIP-chip analysis pipeline, and for processing the much larger data sets that are beginning to arise from multi-million-feature or multiplexed arrays.

There are some simplifications in the current MeDiChI model that could be improved upon in the future. The first of these is that there is no probe-specific hybridization model, (as in Johnson et al., 2006), to account for sequence-related hybridization efficiencies and cross-hybridization. It should be possible to integrate such a model into this procedure, which would enable it to more accurately handle data from array technologies that are more subject to such issues (e.g. Affymetrix 25-mer arrays; Euskirchen et al., 2007). Also, the use of a weighted constrained generalized regression (as in Park and Hastie, 2007) would more accurately handle the intrinsic heteroskedasticity of the intensities, as well as potential nonlinear effects (e.g. saturation) of the microarray.

We also believe that, with simple modifications, the MeDiChI procedure may also aid in the analysis of ‘ChIP-Seq’ data (Johnson et al., 2007). Such data, while of higher effective resolution, will still be limited by the sonicated DNA fragment lengths. The MeDiChI model could be applied to such data by altering the peak profile model to (1) count DNA fragments themselves, rather than their intensities, (2) eliminate the hybridization efficiency term $\Phi(\nu)$, and (3) consider the data to have a resolution of 1 bp. In addition, because of this effective high resolution of ChIP-seq data, we expect that a MeDiChI-based procedure may be able to infer the size of the TF’s footprint ($2 \times \lambda$) on the chromatin.

ACKNOWLEDGEMENTS

We would like to thank Richard Bonneau, Greg Carter, Harri Lähdesmäki, Vesteinn Thorsson, and the anonymous reviewers for helpful comments. This work was supported by NSF Postdoctoral Fellowship DBI 0400598 to M.T.F and the following research grants to N.S.B: NSF (DBI-0640950); DoE (DE-FG02-07ER64327); and NIH (P50 GM076547).

Conflict of Interest: none declared.

REFERENCES


