BIOINFORMATICS

Vol. 23 ISMB/ECCB 2007, pages i80–i86
doi:10.1093/bioinformatics/btm167

A physical model for tiling array analysis

Ho-Ryun Chung*, Dennis Kostka and Martin Vingron

Max-Planck-Institut für molekulare Genetik, Ihnestrasse 63-73, 14195 Berlin, Germany

ABSTRACT

Motivation: Chromatin immunoprecipitation (ChIP) is a powerful experimental approach to identify *in vivo* binding sites of sequence-specific transcription factors (TFs). These experiments are designed to specifically enrich DNA fragments that are bound to the TF. Tiling arrays have become more and more popular for the identification of these DNA fragments. However, many studies showed that only a fraction of the identified DNA fragments contains *bona fide* binding sites for the TF, suggesting that indirect binding mechanisms play a very important role. We explored the possibility that the lack of binding sites can also be explained by problems in identifying ChIP-enriched DNA fragments from the measured intensities.

Results: We derived a physical model that explains some (but not all) variation of the measured probe intensities of Affymetrix tilling arrays. We used the physical model to estimate the probe-specific behavior and corrected for it. Subsequently, we developed a method to identify ChIP-enriched DNA fragments. We termed it physical model for tiling array analysis (PMT). We applied PMT to the data of ChIP-chip experiments interrogating chromosome 21 and 22 of the human genome for binding of the TFs MYC, SP1 and P53. Almost all regions recovered by PMT showed evidence for sequence-specific binding of the TFs.

Contact: chung@molgen.mpg.de

1 INTRODUCTION

Chromatin immunoprecipitation (ChIP) followed by microarray or chip aided identification of enriched DNA fragments (ChIP-chip) is a powerful experimental approach to find *in vivo* transcription factor (TF)-binding sites. The use of tiling arrays to detect enriched DNA fragments has become more and more popular. The main reason for the popularity is that tiling arrays contain probes representing all non-repetitive DNA of chromosome(s) or loci in an unbiased manner, allowing for an unbiased detection of TF-binding sites at high spatial resolution.

Analysis of such experimental data in human revealed that only a fraction of the identified DNA fragments contains *bona fide* binding sites for the TF (e.g. Bieda *et al*., 2006; Cawley *et al*., 2004; Martone *et al*., 2003). The lack of binding sites has been explained by the possibility that the TF may bind to a yet unknown binding motif, or indirectly via another DNA-binding protein, or that the employed antibodies may cross react with other chromatin components (Cawley *et al*., 2004). However, similar experiments in yeast have shown that identified ChIP-enriched DNA fragments can be used to recover motifs by *de novo* motif finding algorithms that resemble known motifs (Harbison *et al*., 2004). Furthermore, it was shown that a significant fraction of genome-wide binding data in yeast could be explained by the occurrence of binding site motifs in the sequences (Roider *et al*., 2007). But the experiments employed different platforms: tiling arrays (human) and microarrays (yeast), where complete intergenic regions of an average length of 1000 base pairs are spotted. It seems possible that the short oligonucleotide probes spotted on tiling arrays are introducing a bias in hybridization intensity (see below). Hence, the failure to detect TF-binding sites may also, at least partially, be contributed for by the methods used to detect enriched DNA fragments from the measured intensities on the tiling array.

The identification of enriched DNA fragments is done by different approaches, which either neglect the probe-specific behavior (e.g. Cawley *et al*., 2004) or explicitly consider it (Ji and Wong, 2005; Johnson *et al*., 2006; Keles *et al*., 2006; Li *et al*., 2005). It seems that the former methods perform not as well as the latter. For example, a direct comparison between the method employed by Cawley *et al*. (2004) and a hidden Markov model approach (Li *et al*., 2005) revealed that while the latter method yielded regions, where the p53-binding motif could be recovered by *de novo* motif finding algorithms, the former failed to do so. It seems that the short oligonucleotide probes on tiling arrays are introducing a bias. More accurate results can be achieved if probe-specific effects are incorporated in the analysis.

Three of the four aforementioned approaches that consider probe-specific effects estimate probe behavior from multiple samples (Ji and Wong, 2005; Keles *et al*., 2006; Li *et al*., 2005) whereas an approach termed model-based analysis of tiling arrays (MAT; Johnson *et al*., 2006) models the probe behavior from a single sample considering only the probe sequence and copy number in the genome. This approach is very similar to the methods used to infer sequence-specific probe effects for gene expression microarrays (Hekstra *et al*., 2003; Wu and Irizarry, 2005) but also includes the effect of the copy number of a probe in the genome.

Our method is similar to the approach of MAT, the major difference being that we do not estimate the probe-specific behavior from the data, but establish it *a priori* using a sequence-dependent physical model, such that we are left with only two parameters, i.e. the slope and the intercept of a linear model. We termed our method physical model for tiling array analysis (PMT). We demonstrate the performance of PMT by applying it to the data of Cawley *et al*. (2004), which performed ChIP-chip experiments to identify binding sites for the TFs MYC SP1 and P53. We found fewer regions than the original study. Some but not all of the loci detected

*To whom correspondence should be addressed.

by PMT overlap with the regions identified in the original study. For MYC as well as SP1 we observed that between 54 and 72% of the detected regions contained the respective binding site consensus. Furthermore, we were able to recover binding site motifs of SP1 and P53, which occur in almost all sequences. The finding that almost all regions detected for SP1 and P53 contain the recovered motifs together with the observation that at least 64% of the MYC-dependent chromosomal regions contain at least one instance of the MYC-MAX consensus motif suggests that the chromatin association of the three TFs can be attributed to sequence-specific binding. Taken together our findings suggest that ChIP-chip experiments can be used to identify mammalian *in vivo* TF-binding sites with high confidence if the probe-specific behavior is explicitly taken into account.

2 METHODS

2.1 Physical modeling of probe-specific intensity

Tiling arrays are designed to contain probes for all non-repetitive sequences of chromosome(s) or loci at a high resolution. Hence, it is not possible to select the probes for their hybridization properties, suggesting that the affinity of the probes for their targets may vary from probe to probe. It has been shown that the probe sequences may be used as a guide to determine the probe-specific behavior of gene expression microarrays (Hekstra *et al.*, 2003; Zhang *et al.*, 2003) but also for tiling arrays used to identify enriched DNA fragments using ChIP experiments (Johnson *et al.*, 2006). These methods use the data to infer the probe-specific intensity bias. While for gene expression arrays typically the data of several experiments is used for this purpose, for tiling arrays a single sample seems to be sufficient (Johnson *et al.*, 2006). This is due to the fact that most probes in a transcription factor (TF) ChIP-chip experiment measure only unspecific binding as TFs usually bind only to a small fraction of the genome.

We take advantage of this property of TF ChIP-chip experiments in order to establish an *a priori* physical model of unspecific binding. We reasoned that the free energy of unspecific binding $\Delta G_u(n)$ of probe p_n is a function of the stability of all DNA duplexes formed by hybridizing the oligonucleotide probe p_n with sequence S_n of length L to the sample DNAs. The ΔG_u of all these duplexes is computed using empirical estimates of the sequence-dependent duplex stability (SantaLucia, 1998) weighted by the expected frequencies of the sample sequences forming these duplexes. The ΔG_u of a duplex starting at base i of the probe and ending at base j is taken to be

$$\exp[-\Delta G_u] = q_{ini}(S_{n,i})p(S_{n,i}) \prod_{k=i+1}^{j} \left[q_{sym}(S_{n,k}|S_{n,k-1})p(S_{n,k}|S_{n,k-1})\right] q_{ini}(S_{n,j}) \quad (1)$$

where $S_{n,i}$ denotes the base of sequence S_n at position i. The $q_{sym} = \exp[-\Delta G_{sym}]$ terms are dependent on the base at position k given the base at position $k-1$, effectively accounting for both hydrogen bonds and stacking energy. The $q_{ini} = \exp[-\Delta G_{ini}]$ terms account for the base pairs having no (single base pair) or only one (duplexes longer than a single base pair) neighboring base pair. $p(S_{n,k})$ is the frequency of the base $S_{n,k}$ in the human genome (NCBI build 35) counting both the sense and anti-sense strand, while $p(S_{n,k}|S_{n,k-1})$ is the frequency of base $S_{n,k}$ given the base $S_{n,k-1}$. We consider all possible duplexes ranging from a single base pair to duplexes formed by all L bases of the probe. We employ a dynamic programming technique to calculate the ΔG_u using following recursion relations:

$$\begin{aligned}\hat{Q}(i) &= q_{ini}(S_{n,i})p(S_{n,i}) + \\ &\quad q_{sym}(S_{n,i}|S_{n,i-1})p(S_{n,i}|S_{n,i-1})\hat{Q}(i-1) \\ Q(i) &= Q(i-1) + \hat{Q}(i)q_{ini}(S_{n,i}).\end{aligned} \quad (2)$$

The recursion starts at position 2, with the initial conditions $\hat{Q}(i=1) = q_{ini}(S_{n,1})p(S_{n,1})$ and $Q(i=1) = \hat{Q}(1)q_{ini}(S_{n,1})$, and ends at position L. The ΔG_u for a probe is calculated by

$$\Delta G_u = -\log(Q(L)). \quad (3)$$

We assume that any of these duplexes forms independent of the others and that there is no competition between the duplexes.

2.2 A simple model for probe intensities

Here we use the probe-specific free energy for unspecific binding, ΔG_u, to model the intensity signal we expect from each probe on an array. With this approach we account for sequence-specific signal components in a physically motivated manner. Dealing with ChIP-chip data we expect mainly 'background' or unspecific hybridization to contribute to the measured intensity. Further on we assume variation in the signal due to optical noise to be small compared to that arising from other sources (Wu and Irizarry, 2005). Therefore we subtract an estimated constant optical background of $\hat{O} = \min(\text{PM}) + 1$ from each array (Wu and Irizarry, 2005).

The resulting (background corrected) intensity values correlate well with what we expect to see for unspecific binding (see Results section for details). These observations motivated us to work with the following simple model. We assume the signal intensity of each PM to be a random variable composed of two independent contributions, namely an optical background plus a non-specific binding signal: $\text{PM} = B_{\text{PM}} + O$. For O we use the plug-in estimate $\hat{O}$ discussed above. The background hybridization we assume to depend linearly on the free energy for unspecific binding:

$$\log(B_{\text{PM}}) = \alpha + \beta\Delta G_u + \epsilon. \quad (4)$$

Here ϵ is a zero mean symmetric random variable accounting for random deviations from our model. Note that we summarize all 'non-signal' contributions into B_{PM}. That is, it may contain non-specific binding as well as additional 'background components'. This, and that we expect comparably few probes with a true signal, allows us to fit the two free parameters α and β in Equation (4) using simple linear regression. Probes with a larger intensity than predicted are then assumed to carry evidence for TF binding. As a score of evidence we use $s_i = (\log(\text{PM}_i - \hat{O}) - \log(\hat{B}_{\text{PM}_i}))/\hat{\sigma}_i$, where $\hat{\sigma}_i$ is the SD of scores in a bin of predicted intensities containing $\hat{B}_{\text{PM}_i}$ (Johnson *et al.*, 2006).

To accumulate evidence of contiguous positions we slide a window along the chromosome. We average all scores in a 500 bp window and multiply the result with the square root of contributing probes (similar to Johnson *et al.*, 2006). Finally, windows exceeding an (accumulated) score cutoff are candidates for TF binding. Empirical P-values can be derived by postulating that the negative part of the score distribution reflects a suitable null hypothesis. This concept may also be employed to calculate 'regional'' false discovery rate (Johnson *et al.*, 2006).

In all reported analyses we set arbitrarily a score cutoff of five. We merged all overlapping windows and report the scores of the highest scoring window. These are the chromosomal regions that were subsequently analyzed.

2.3 Input data

The data reported by Cawley *et al.* (2004) was downloaded from http://transcriptome.affymetrix.com/publication/tfbs. Each experiment and

the controls were conducted on three chips, referred to as A, B and C, respectively. In total there were four experiments, one for MYC and SP1 and two for P53 (employing two different antibodies, hereafter referred to as P53-DO1 and P53-FL), and two matched controls, one using the total INPUT DNA and the other using an antibody against bacterial GST. Therefore there are eight possible combinations of experiments and matched controls:

- MYC compared with total INPUT DNA (INPUT)
- MYC compared with anti-GST DNA (GST)
- SP1 compared with INPUT
- SP1 compared with GST
- P53-DO1 antibody (P53-DO1) compared with INPUT
- P53-DO1 compared with GST
- P53-FL antibody (P53-FL) compared with INPUT
- P53-FL compared with GST.

Human genome sequences (NCBI build 35) were downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/hg17/chromosomes/. We remapped the probe sequences to the genome using PrimerMatch (http://ftp.bioinformatics.org/pub/PrimerMatch/) and kept unique probes that map to chromosome 21 or 22.

2.4 Search for consensus binding sites

We searched the candidate regions for binding site consensus motifs using the patterns `CA[CT]G[TC]G` for MYC, `GG[GT]G[CT]GGG` for SP1 and `X[0-14N]X` with `X = [AG][AG][AG]C[AT][AT]G[CT][CT][CT]` for P53. As the candidate sequences were extracted from already repeat masked whole genome sequences (indicated by small letters), we did not perform any additional repeat masking. We counted the number of sequences with at least one instance of the pattern.

2.5 *De novo* motif discovery

The candidate regions were used as input for the motif finding program MEME version 3.0 (Bailey and Elkan, 1994). We searched for ten motifs with a minimal width of six and a maximal width of 20 bp, only for P53 we increased the minimal width to 10 bp. In order to check whether the motifs are also over-represented compared to the human background, we used a program called CLOVER (Frith *et al*., 2004). We converted the MEME motifs into count matrices. These count matrices and, as a background control, the sequences of human chromosome 21 and 22 (NCBI build 35) served as input for CLOVER. Binding site motifs found to be over-represented compared to this background control were manually inspected for resemblance to the known motifs reported by JASPAR (Sandelin *et al*., 2004). Sequence logo representations were prepared using weblogo (Crooks *et al*., 2004).

3 RESULTS

We applied PMT to the MYC, SP1 and P53 chromatin immunoprecipitation-chip (ChIP-chip) data (Cawley *et al*., 2004). The data was derived from experiments where labeled ChIP-enriched DNA fragments were hybridized to Affymetrix tiling arrays covering human chromosomes 21 and 22. As outlined in the Methods section and summarized in Table 1, we analyzed the eight possible combinations of four experiments and two different controls, i.e. DNA fragments derived from a ChIP-experiment employing an antibody against bacterial GST (GST) and total input DNA (INPUT).

Table 1. PMT detected loci

Experiment	Number of loci	% overlap with original study	% loci with ≥1 motif
MYC versus GST	66	97 (64)	72
MYC versus INPUT	78	95 (74)	64
SP1 versus GST	37	97 (36)	54
SP1 versus INPUT	53	100 (53)	62
P53-DO1 versus GST	23	65 (15)	0
P53-DO1 versus INPUT	24	58 (14)	0
P53-FL versus GST	12	92 (11)	0
P53-FL versus INPUT	16	69 (12)	0

We counted the number of PMT detected regions also found by the original study and report them in brackets.

3.1 Prediction of log-intensities by ΔG_u

The analysis of ChIP-chip data revealed that only a fraction of the identified chromosomal regions contained *bona fide* TF-binding sites (e.g. Bieda *et al*., 2006; Cawley *et al*., 2004; Martone *et al*., 2003). It seems to be possible that this apparent lack of binding sites is due to binding of the TF to a yet unknown binding motif, indirect binding via another chromatin associated protein or cross-reacting antibodies. However, yet another explanation may be that the methods to detect ChIP-enriched DNA fragments are not specific enough. The analysis of gene expression microarrays revealed that the probe sequence itself introduces a bias in the measured intensities (Wu and Irizarry, 2005).

A typical ChIP-chip experiment interrogates the genomic positions of binding events of sequence-specific TFs. We expect therefore that only a small fraction of the genome will be enriched after such an experiment. It is straightforward to assume that the vast majority of probes in ChIP-chip experiments will measure only unspecific binding. Owing to this expectation it has been shown that the measured intensities can be used to infer the probe-specific effect even from a single sample (Johnson *et al*., 2006). We reasoned that the probe-specific effect can be estimated *a priori* using a physical model of unspecific binding. Briefly, we calculate the free energy change ΔG_u of unspecific binding by taking into account every possible stretch of base pairing (duplex) between the probe and all possible sequences in the human genome. We weight every possible duplex by the frequency of a reverse complementary sequence in the human genome (see Methods section for details).

We tested our simple physical model by checking for correlation between the ΔG_us and the (background corrected; see Methods section) perfect match intensities on a logarithmic scale. The Pearson correlation coefficient r ranges from 0.45 to 0.73 (median: 0.66) on the different microarrays (see Fig. 1 for an example). ΔG_us can capture between 20 and 53% (median: 44%) of the variance observed in the data.

Encouraged by these results we developed a method that allows for detection of ChIP-enriched DNA fragments based

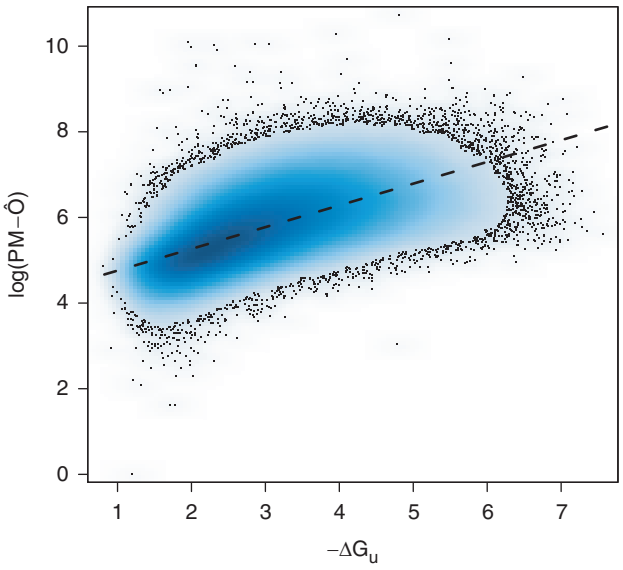

Fig. 1. Smoothed out scatter plot of PM against $-\Delta G_u$. The PM values are from a C-chip interrogating P53-FL. In this case the Pearson correlation is $r = 0.64$ and ΔG_u captures a general trend in the *PM* signal.

on our physical model of probe-specific behavior (see Methods section for details), hence we called it PMT.

3.2 Comparison between PMT and original results

PMT detected between 66 and 78 (41 overlapping) distinct loci for MYC, between 37 and 53 (28 overlapping) for SP1 and finally between 12 and 24 (11 overlapping) for p53 depending on the controls and/or the antibody (see Table 1 for details). Thus, compared with the original study, where 756 loci for MYC, 353 for SP1 and 48 for P53 were reported (Cawley *et al.*, 2004), we found substantially fewer regions. However, most but not all of the chromosomal regions detected by PMT (between 58 and 100%, see Table 1 for details) have also been found by the original study. This observation suggests that PMT is more specific than the method used in the original study.

We expect that ChIP-enriched DNA fragments should be characterized by the presence of sequence motifs that allow for the chromatin association of the TF in a sequence-dependent manner. In principle, chromatin association can be achieved either directly, or indirectly via another DNA-binding chromatin component. A necessary condition for direct binding is the presence of binding sites. As the binding site consensus sequences are known for all three TFs, we checked whether they occur at all in the identified regions. The results are summarized in Table 1. For MYC, we found that the consensus sequence is present in 72% (48 of 66) of the regions detected by the comparison of MYC with the GST-control and 64% (50 of 78) of the regions employing the INPUT-control. For SP1 we found the consensus sequence in 54% (20 of 37) and 62% (33 of 53) of the loci detected by the comparison with the GST- and INPUT-controls, respectively. Only for P53 we found not a single instance of the consensus motif in any of the four possible combinations of antibody and control. The original study reported 33, 22 and 2% of the sites contained the consensus motifs of MYC, SP1 and P53 (Cawley *et al.*, 2004), indicating that the enrichment of the identified chromosomal regions can only in part be explained by sequence-specific binding of the TFs. Our results, however, suggest that, with the exception of P53, the enrichment of the detected regions can be explained much better by a sequence-dependent association of the TFs. It is also possible that we found more instances of the consensus motifs because we applied a much more stringent threshold. Thus, we re-analyzed the data by Cawley *et al.* (2004) by choosing *P*-value cut-offs that result in comparable numbers of loci. Within these newly generated datasets, we identified comparable numbers of consensus motifs for MYC and SP1 (data not shown). This observation suggests that the *P*-value cut-off chosen in the original study was not stringent enough. However, we would like to note that PMT also found regions that were not detected in the original study, indicating that probe-specific properties can influence the analysis done in the original study (see below).

3.3 *De novo* motif discovery

The analysis reported above has the caveat that binding site consensus sequences are usually defined by *in vitro* experiments. It is possible that the *in vivo* binding sites differ significantly from the *in vitro* ones, as it may be the case for P53. Moreover, experiments do not have to necessarily involve TFs whose binding site motif is known in advance. In such cases it is desirable to be able to recover a small number of candidate motifs, which can be tested experimentally. As we know the binding site motifs for MYC, SP1 and P53 *a priori*, we can now assess whether *de novo* motif finding algorithms can recover motifs that resemble the known ones. Motif finding algorithms applied to ChIP-chip regions often recover motifs that are not over-represented in the regions compared to the genomic background. Their significance is only due to the 'oddness' of the motif compared to the background model used in the motif finding algorithm. In order to eliminate such motifs, we ran CLOVER (Frith *et al.*, 2004) to check every motif recovered by MEME (Bailey and Elkan, 1994) whether it is over-represented in the identified DNA sequences compared to the human genomic background. We discarded every motif that failed to be significantly over-represented in such a comparison.

For MYC we were able to recover a single motif, namely from the comparison to the GST control. As MYC typically binds as a heterodimer with MAX, we compared the motif to the MYC-MAX motif as reported by JASPAR (Sandelin *et al.*, 2004) and found no significant overlap between the two motifs.

For SP1, we recovered four motifs for the regions detected in comparison with the GST-control and also four motifs for regions using the INPUT control. In both cases, we found a motif (with the second highest CLOVER score) that contained multiple copies of the known SP1-binding motif as reported by JASPAR (Fig. 2 a–c). Significantly, both motifs occur in all

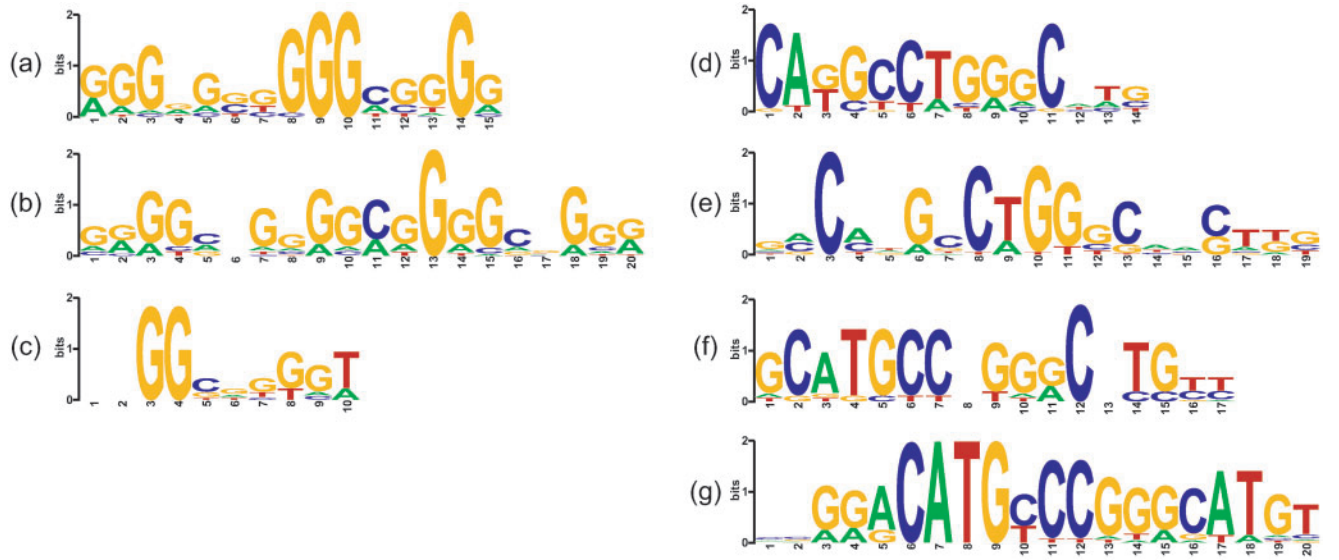

Fig. 2. Sequence logo representation of identified motifs. (**a**) SP1 motif identified by MEME in chromosomal regions detected in a comparison between SP1 and GST-control; (**b**) SP1 motif identified by MEME in chromosomal regions detected in a comparison between SP1 and INPUT-control; (**c**) SP1 motif as reported by JASPAR; (**d**) P53 motif identified by MEME in chromosomal regions detected in a comparison between P53-DO1 and GST-control; (**e**) P53 motif identified by MEME in chromosomal regions detected in a comparison between P53-DO1 and INPUT-control; (**f**) P53 motif identified by MEME in chromosomal regions detected in a comparison between P53-FL and GST-control and (**g**) P53 motif as reported by JASPAR.

37 and 53 regions identified by comparing against the GST- and INPUT-control, respectively. Furthermore, we observe that the SP1 motif as reported by JASPAR is not over-represented in the regions detected by PMT compared to the genomic background. This finding suggests that multiple copies of the binding sites are necessary for efficient and specific binding of SP1.

Finally, the P53-dependent chromosomal regions yielded between three and ten motifs depending on the antibody and the controls. Except for the combination of the P53-FL antibody with the INPUT-control we recovered motifs (position two in the CLOVER score list for P53-DO1 and P53-FL compared to GST, and the highest scoring motif for P53-DO1 against INPUT control) that strongly resemble the P53-binding motif as reported by JASPAR (Fig. 2 d–g). We note that the regions detected by the P53-FL antibody and the INPUT-control contained a motif that resembles the P53-binding site but was not found to be over-represented using the human genome as control. For the P53-DO1 datasets, we found the motifs in all 23 regions detected by comparison with GST and in all 24 regions using the INPUT-control. The motif recovered from a comparison between the P53-FL dataset and GST-control occurs in 11 out of 12 sequences. P53 is known to bind to two palindromic half sites that are separated typically by zero to 13 bp (el Deiry *et al*., 1992). Thus, it is not surprising that we find for all three reported motifs that the first half site is always stronger than the second one (Fig. 2 d–f), owing to the fact that the second half site may not be directly adjacent to the first one.

We were able to recover binding site motifs for SP1 and P53 using a *de novo* motif finding algorithm. These resemble the known binding site motifs reported by JASPAR and were in all cases among the two most significantly over-represented motifs reported by CLOVER. In the original study only the recovery of the SP1 motif was reported (Cawley *et al*., 2004). We speculate that the failure to detect the binding site motif of P53 in the original study has two reasons: (i) some of the detected regions were false positives and (ii) the original study missed some real positives. For example, we re-analyzed the dataset by comparing the P53-DO1 antibody to the GST control. We applied a much more stringent *P*-value cut-off that was chosen to include regions called with the 23 lowest *P*-values. This resulted in 24 distinct loci, a number that is comparable to the 23 loci detected by PMT. For these 24 sequences we recovered a motif that strongly resembles the P53-binding motif as reported by JASPAR (data not shown). However, this motif was present only in 13 sequences. Moreover, we found that nine of these 13 sequences correspond to loci identified also by PMT. These results are in support of the argument that the identified regions of the original study contained false positives as well as false negatives.

The finding that the recovered SP1- and P53-binding site motifs occur in almost all detected DNA fragments suggests that the chromatin association of these two TFs is primarily due to direct binding of the proteins to DNA. Given the success for SP1 and P53, it seems rather unlikely that we failed to identify the binding site motif for the MYC-MAX heterodimer due to problems in finding MYC-associated chromosomal regions. It seems to be much more plausible that this failure is due to either indirect binding of MYC via another DNA binding protein and/or unspecific binding of the MYC-antibody to other proteins. Yet another explanation may be

that MYC-MAX binding sites are not over-represented in the sequences possibly because they are too short.

Given that we found the MYC-MAX consensus sequence (see above) in at least 64% of the sequences suggests that direct binding of MYC can account for the chromatin association in the majority of cases. Still, there are DNA fragments that cannot be explained by direct binding of MYC- to its consensus motif, indicating that chromatin association of MYC is also possible by indirect binding via other chromatin components. In a recent study using ChIP followed by pair-end ditag sequencing of the ChIP-enriched DNA fragments it was shown that up to 40% of the highly significant ChIP-enriched chromosomal regions do not contain any identifiable MYC-binding site (Zeller *et al*., 2006). It may still be possible that the lack of binding sites is due to cross reactions of the MYC- antibody. But given the high percentage of sequences containing a MYC binding site, we favor instead the argument that MYC can also bind indirectly via other chromatin components.

4 DISCUSSION

We developed and implemented a method, referred to as PMT that allows for the recovery of ChIP-enriched DNA fragments. It is based on a physical model for unspecific binding that has been shown to capture some of the probe-specific behavior.

The physical model presented here can, in principle, be applied to other microarray platforms, such as gene expression arrays or tiling arrays used to delineate transcription units. However, the probes on these microarrays do not primarily measure unspecific binding, such that specific binding may become an issue. Given that our physical model seems to capture unspecific binding very well it may be of interest to develop a physical model for specific binding and to combine both probe characteristics into a single model. This unified model may help to normalize the measured hybridization intensities, which in turn is required for the comparison between microarrays and experiments.

The approach of PMT resembles MAT (Johnson *et al*., 2006), with the major difference being that PMT does not fit the probe-specific behavior from data but establishes it *a priori* using a physical model. If we compare the PMT results with the results obtained by MAT runs, we find a large overlap (data not shown). Thus, it seems that our physical model can capture the probe-specific effects in a comparable manner to the parameter-rich model of MAT.

We showed that the detected chromosomal regions are likely to be enriched due to sequence-specific binding of the TF in question. One merit of ChIP-chip experiments is the identification of *in vivo* binding sites for TFs. We recovered in two out of three cases the known binding site motifs among the two most significantly over-represented motifs reported by CLOVER. The recovered binding site motifs of SP1 suggest that efficient and specific binding of SP1 is contingent on the presence of multiple binding sites. The recovered motifs are much more specific, because they are longer than the SP1 motif reported by JASPAR. Thus, it is possible that the recovered motifs for SP1 can be used to search for SP1 binding sites in the human genome sequence and in turn may facilitate the identification of SP1-responsive genes *in silico*. The same is true for P53: the identified motifs resemble the known P53-binding site motifs but are not identical. For instance the positions six and nine of the JASPAR P53 motif require C and G at these positions, while the recovered motifs are less strict (compare for example Fig. 2 d and g). A similar finding was reported by Li *et al*. (2005). They were able to recover the P53 motif from 43 high-confidence regions. However, the motif occurred only in 40% of their regions, while the motifs detected in PMT identified loci occur in almost all (92–100%) sequences. We showed that a similar result as reported by Li *et al*. (2005) can be obtained if one restricts the original hit list of Cawley *et al*. (2004) to the most significant sequences. Both, Li *et al*. (2005) as well as Cawley *et al*. (2004) used quantile normalization. Hence, it appears that the results of *de novo* motif finding algorithms depend on the probe-level normalization strategy.

Taken together, we showed evidence that PMT is able to faithfully recover ChIP-enriched DNA fragments from ChIP-chip experiments. The analysis of the regions showed that most but not all detected ChIP-enriched DNA fragments can be explained by sequence-specific binding of the TF in question. Thus, we would like to conclude that ChIP-chip experiments can be used to detect *in vivo* binding sites of TFs.

Conflict of Interest: none declared.

REFERENCES

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.

Bieda,M. *et al*. (2006) Unbiased location analysis of e2f1-binding sites suggests a widespread role for e2f1 in the human genome. *Genome Res.*, **16**, 595–605.

Cawley,S. *et al*. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding rnas. *Cell*, **116**, 499–509.

Crooks,G.E. *et al*. (2004) WebLogo: a sequence logo generator. *Genome Res*, **14**, 1188–90.

elDeiry,W. *et al*. (1992) Definition of a consensus binding site for p53. *Nat Genet.*, **1**, 45–9.

Frith,M.C. *et al*. (2004) Detection of functional dna motifs via statistical over-representation. *Nucleic Acids Res.*, **32**, 1372–1381.

Harbison,C.T. *et al*. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Hekstra,D. *et al*. (2003) Absolute mrna concentrations from sequence-specific calibration of oligonucleotide arrays. *Nucleic Acids Res.*, **31**, 1962–1968.

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

Johnson,W.E. *et al*. (2006) Model-based analysis of tiling-arrays for chip-chip. *Proc. Natl Acad. Sci. USA*, **103**, 12457–12462.

Keles,S. *et al*. (2006) Multiple testing methods for chip-chip high density oligonucleotide array data. *J. Comput. Biol.*, **13**, 579–613.

Li,W. *et al*. (2005) A hidden markov model for analyzing chip-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics*, **21** (Suppl. 1), i274–i282.

Martone,R. *et al*. (2003) Distribution of nf-kappab-binding sites across human chromosome 22. *Proc. Natl Acad. Sci. USA*, **100**, 12247–12252.

Roider,H.G. *et al*. (2007) Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, **23**, 134–41.

Sandelin,A. *et al*. (2004) Jaspar: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*., **32**,D91–D94.

SantaLucia,J. (1998) A unified view of polymer, dumbbell, and oligonucleotide dna nearest-neighbor thermodynamics. *Proc*. *Natl Acad. Sci. USA*, **95**, 1460–1465.

Wu,Z. and Irizarry,R.A. (2005) Stochastic models inspired by hybridization theory for short oligonucleotide arrays. *J*. *Comput*. *Biol*., **12**, 882–893.

Zeller,K.I. *et al*. (2006) Global mapping of c-myc binding sites and target gene networks in human b cells. *Proc*. *Natl Acad. Sci. USA*., **103**, 17834–17839.

Zhang,L. *et al*. (2003) A model of molecular interactions on short oligonucleotide microarrays. *Nat*. *Biotechnol*., **21**, 818–821.