BIOINFORMATICS

Vol. 21 Suppl. 1 2005, pages i274–i282
doi:10.1093/bioinformatics/bti1046

A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences

*Wei Li, Clifford A. Meyer and X. Shirley Liu**

Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Harvard School of Public Health, Boston, MA 02115, USA

Received on January 15, 2005; accepted on March 27, 2005

ABSTRACT

Motivation: Transcription factors (TFs) regulate gene expression by recognizing and binding to specific regulatory regions on the genome, which in higher eukaryotes can occur far away from the regulated genes. Recently, Affymetrix developed the high-density oligonucleotide arrays that tile all the non-repetitive sequences of the human genome at 35 bp resolution. This new array platform allows for the unbiased mapping of *in vivo* TF binding sequences (TFBSs) using Chromatin Immuno-Precipitation followed by microarray experiments (ChIP-chip). The massive dataset generated from these experiments pose great challenges for data analysis.

Results: We developed a fast, scalable and sensitive method to extract TFBSs from ChIP-chip experiments on genome tiling arrays. Our method takes advantage of tiling array data from many experiments to normalize and model the behavior of each individual probe, and identifies TFBSs using a hidden Markov model (HMM). When applied to the data of p53 ChIP-chip experiments from an earlier study, our method discovered many new high confidence p53 targets including all the regions verified by quantitative PCR . Using a *de novo* motif finding algorithm MDscan, we also recovered the p53 motif from our HMM identified p53 target regions. Furthermore, we found substantial p53 motif enrichment in these regions comparing with both genomic background and the TFBSs identified earlier. Several of the newly identified p53 TFBSs are in the promoter region of known genes or associated with previously characterized p53-responsive genes.

Contact: xsliu@jimmy.harvard.edu

Supplementary information: Available at the following URL http://genome.dfci.harvard.edu/∼xsliu/HMMTiling/index.html

1 INTRODUCTION

One of the most important characteristics of gene regulation is the interaction between transcription factors (TFs) and *cis*-regulatory elements. Although microarrays have been widely used to understand gene-expression regulation, they do not provide information on which TFs directly regulate which genes and their interaction mechanism. In recent years, ChIP-chip has become a popular technique for studying the genome-wide location of *in vivo* TF–DNA interactions. ChIP-chip was first successfully adopted in yeast to identify the regulatory targets of individual TFs (Ren *et al*., 2000; Lieb *et al*., 2001) and to study the entire transcriptional regulatory networks (Harbison *et al*., 2004). Promoter arrays were used in the yeast ChIP-chip experiments, with a cDNA probe for each intergenic sequence. Since higher eukaryotes, especially human, have long intergenic sequences, the promoter arrays for higher eukaryotes usually contain probes just for the proximal promoters of annotated genes (Li *et al*., 2003; Odom *et al*., 2004). Unfortunately, TFBSs in higher eukaryotes can occur upstream, downstream, close to or far away from the regulated genes, or even in the introns of the genes. Thus, proximal promoter arrays may not accurately capture all the ChIP-enriched DNA. Although CpG island (Wells *et al*., 2003) and continuous genomic PCR fragment (Euskirchen *et al*., 2004) arrays have been used to address this problem, their probe densities and genome coverage are still not satisfactory.

Recently, Affymetrix developed the high-density oligonucleotide arrays that tile all non-repetitive sequences of the human genome. Current tiling of chromosomes 21 and 22 (Kapranov *et al*., 2002) are available and the whole genome tiling arrays are under development. These arrays have one probe pair, a perfect match (PM) probe and a mismatch (MM) probe both 25 bases long, for every non-overlapping 35 bp region in the genome. They provide the platform for the unbiased mapping of *in vivo* TF binding sequences (TFBSs) using ChIP-chip, but the massive amount of data (∼1 million probe pairs for chromosomes 21 and 22) generated from them pose great challenges for data analysis.

Cawley *et al*. (2004) conducted p53 ChIP-chip experiments on chromosomes 21 and 22 tiling arrays. Six duplicated experiments were performed on each of two p53 antibodies: FL (full length) and DO1 (N-terminal epitope), as well as two control experiments: Input (genomic input DNA without ChIP)

*To whom correspondence should be addressed.

and GST (antibody to bacterial GST). The GTRANS software provided by Affymetrix (http://www.affymetrix.com/support/developer/downloads/TilingArrayTools/index.affx) was used to predict p53 binding sequences for each antibody against each control (FL–Input, FL–GST, DO1–Input and DO1–GST). GTRANS requires at least three replicate TF ChIPs and three replicate controls to find the ChIP-enriched regions. To check whether a probe x is ChIP-enriched, all probes within 500 bp from x in all ChIP and control experiments are considered. GTRANS uses the non-parametric Mann–Whitney U-test (equivalent to Wilcoxon rank sum test) to rank all the probe pairs by their log2(max(PM − MM),1)) values, and checks whether the sum of ranks of all probe pairs in the ChIPs are significantly higher than that in the controls. The two datasets with the same antibody were merged together to form a non-redundant set (48 FL sites reported in the paper and 103 DO1 sites downloadable from the Supplementary website). A total of 17 and 0% of the identified p53 binding sequences were located within 1 kb of CpG islands and 5′ exons, respectively, indicating that only a small fraction of p53 sites would have been discovered by using CpG island arrays or proximal promoter arrays. When trying to find putative TF binding motifs from the 48 p53-FL sites, Cawley *et al*. (2004) failed to identify the p53 binding motif.

Although the Mann–Whitney U-test assumes no specific probability distribution of the data, it cannot identify enriched regions with small P-values without a long enough window and enough replicates. We developed a new algorithm and applied it to the same p53-FL dataset. We first normalized the data across all datasets published in Cawley *et al*. (2004), removed all the repetitive probe measurements and estimated the behavior of each probe. Then a two-hidden-state (ChIP-enriched state and non-enriched state) hidden Markov model (HMM) (Rabiner, 1989) was implemented to estimate the probability of ChIP enrichment at each probe location. We identified many potential p53 binding sequences, including all the quantitative PCR verified targets. From the high confidence regions, a p53 binding motif was successfully recovered using MDscan (Liu *et al*., 2002). Furthermore, we found substantial p53 motif enrichment in our HMM identified regions compared with both genomic background and TFBSs identified by Cawley *et al*. (2004). The entire analysis is summarized in Figure 1.

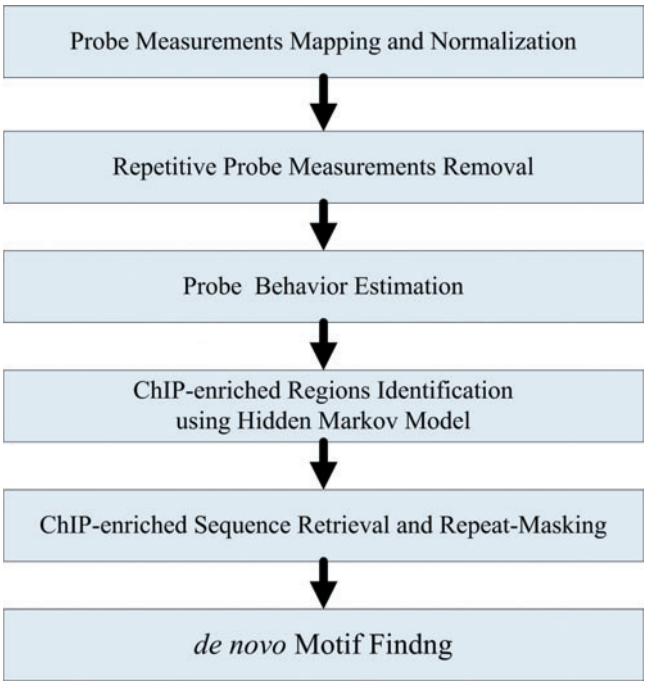

Fig. 1. Strategy diagram for analyzing ChIP-chip experiments on genome tiling arrays.

2 METHODS

2.1 Dataset

Cawley *et al*. (2004) conducted ChIP-chip experiments for four human TFs in two cell lines (p53-FL and p53-DO1 in HCT1116, cMyc and Sp1 in Jurkat) as well as two control experiments (input and GST) in each cell line. Each experiment contains two biological replicates and three technical replicates, generating a total of 48 datasets (samples). Each sample was hybridized to the Affymetrix chromosomes 21 and 22 tiling arrays (three arrays, A, B and C). All 48 datasets were downloaded from http://transcriptome.affymetrix.com/publication/tfbs/. The (PM−MM) value was recorded for each probe pair as a new probe value. To be able to compare probe values between samples (Workman *et al*., 2002), we conducted quantile normalization (Bolstad *et al*., 2003) on each array (A, B and C) separately to make the probe value distribution the same across all samples. We then mapped the probe values from the three arrays to the April 2003 version of the human genome according to the Affymetrix NCBIv33 GTRANS Library.

2.2 Repetitive probe measurements

Owing to the array coverage and genome sequence similarity, occasionally N probes with the same 25mer sequence might be spotted at different locations on the chips. At the mean time, a 25mer might map to M different genomic locations. Therefore, a 25mer sequence might have up to $N \times M$ repetitive probe measurements (Fig. 2). Since ChIP-enriched regions are normally <1 kb, short-range repetitive probe measurements (<1 kb of each other) with high values will enhance the chance of the region to be falsely predicated as ChIP-enriched, and thus should be removed. In contrast, long-range repetitive probe measurements (>1 kb of each other) should be kept for they may represent valid targets (although cross-hybridization issue should be addressed in future study). We decided to employ the following two-stage procedure to filter out short-range repetitive probe measurements prior to

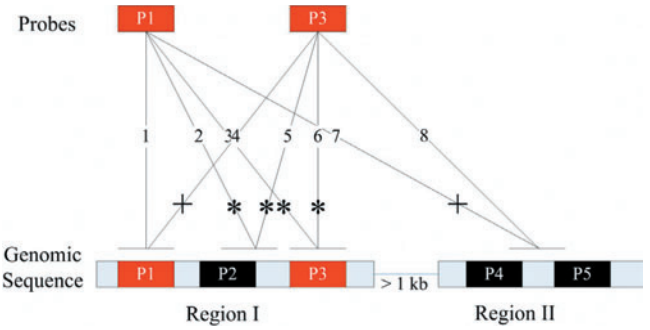

Fig. 2. Repetitive probe measurements. Probes P1 and P3 ($N = 2$) have the same 25mer sequences, each maps to four genomic location ($M = 4$), thus generating eight ($N \times M$) repetitive probe measurements. Repetitive probe measurements are considered short-range if they are within 1 kb (within Regions I and II) or long-range otherwise (between Regions I and II). In the two-stage procedure to filter out short-range duplications, four measurements (2, 3, 5, 6; indicated as *) are filtered out at Stage I, two measurements (4, 7; indicated as +) are removed at Stage II.

downstream analysis: (1) Ensure only one measurement per probe within 1 kb of the genome. For each probe mapping to M multiple genomic locations, from the beginning of each chromosome, we deleted the latter occurrence if the genomic distance between two consecutive occurrences is <1 kb of each other. This process was iterated until all the probe occurrences had been considered. (2) Ensure only one probe measurement for each genomic position. We have three (A, B and C) different kinds of tiling arrays in our current datasets. The N multiple probes on the same genomic position might be from different arrays. Since we quantile normalized each array separately, there could be a huge array bias if we take the average of all the duplicated probe measurements. Therefore, we randomly chose one probe and discarded the others.

2.3 Probe behavior estimation

ChIP-chip experiments enrich only a small portion of DNA bound by the TF of interest, and control experiments capture only baseline non-specific binding. Therefore, most of the probe measurements across all datasets are not actually hybridized with ChIP-enriched DNA and could be used to estimate the baseline probe behavior. We randomly selected 5000 probes from the ~1 million probe dataset and conducted Shapiro–Wilk normality test on the behavior (PM−MM) of each probe across the 48 datasets. The majority (~80%) of these probes have the Shapiro–Wilk normality P-value $>$ 0.01, accepting the null hypothesis that the distribution of probe behavior across the 48 datasets is normal. The remaining 20% probes that are not normal could be caused by ChIP or non-specific enrichment in some of the 48 datasets. We estimated the baseline probe behavior using 36 of the 48 datasets, including all of the control datasets and 99.5% of probes from the non-p53 ChIP experiments (excluding the top 0.5% percentile ~5000 probes). The 12 p53 ChIP datasets (FL and DO1) were excluded from the parameter estimation because their ChIP enrichment signal may skew the baseline estimate. The behavior of each probe i is modeled as a normal distribution $N(\mu_i, \sigma_i^2)$, where μ_i and σ_i are the mean and standard deviation of probe i values of the 36 tiling array datasets.

2.4 Hidden Markov model

We designed a two-hidden-state (ChIP-enriched state and non-enriched state) HMM to estimate the probability of enrichment at each probe location i. Given J potential binding sites (assumed to be ~300 along chromosomes 21 and 22 for p53) along chromosomes covered by K total probes, the HMM is characterized by the following:

(1) Initial probabilities: J/K for ChIP-enriched state, $1 - $ J/K for non-enriched state.

(2) Transition probabilities: J/K for transition to a different state, $1 - $ J/K for staying in the same state.

(3) Emission probability distribution of probe i in single dataset: $N(\mu_i + 2\sigma_i, (1.5\sigma_i)^2)$ for ChIP-enriched state, $N(\mu_i, \sigma_i^2)$ for non-enriched state. The parameters are based on the results on the Affymetrix SNP arrays (Lieberfarb *et al.*, 2003).

(4) A probe i, with (PM−MM) value p_i, is defined as an outlier if its Z-value is >3 or <-2.5. We reassigned the Z-value of each outlier probe as 3 if $Z > 3$ and -2.5 if $Z < -2.5$.

(5) If two adjacent probes are farther apart than 500 bp in the genome (usually due to a long repeat sequence between the two probes), in the forward and backward procedure, the enriched and non-enriched state probabilities of the latter probe are reset to the initial probabilities.

To combine the results from different replicates, in either the ChIP or control group, the emission probabilities of all available replicates for each probe were averaged as the emission probability on this probe. The forward and backward algorithms of HMM (Rabiner, 1989) were used to calculate the probabilities of a probe being in ChIP-enriched versus non-enriched states. A log-odds enrichment value was given to each probe representing the ratio of being ChIP-enriched against non-enriched state. The natural cutoff for the ChIP-enriched state is 0, which indicates equal ChIP-enriched and non-enriched probabilities.

The ChIP-enriched region is defined as at least two probes with log-odds enrichment value >0 in ChIP and at least one probe with log-odds enrichment value <-15 in the control. The average enrichment value of all the probes in a ChIP-enriched region is used as a summary enrichment score for the entire region.

2.5 Sequence retrieval and repeat masking

Genomic sequences of the HMM identified ChIP-enriched regions were retrieved from the repeat-masked April 2003 version of the human genome at UCSC Genome Browser (http://genome.ucsc.edu/). We further masked out all the tandem repeats identified by Tandem repeats finder (Benson, 1999) to facilitate downstream *de novo* motif finding. The resulting sequences are defined as fully-repeat-masked sequences.

2.6 Motif finding and enrichment scoring

We ranked all the fully-repeat-masked ChIP-enriched sequences by the summary enrichment score and applied MDscan (Liu *et al*., 2002) to find the putative p53 binding motif. For a motif of width w, MDscan first enumerates each $w$mer in the highest ranking sequences, and collects other $w$mers similar to it in these sequences to construct a candidate motif, represented as a probability matrix. A semi-Bayesian scoring function is used to remove low scoring candidate motifs and refine the rest by checking all $w$mers in all the ChIP-enriched sequences. If several final motifs share the same consensus, only the motif with the highest score is kept (Conlon *et al*., 2003).

We determined how well a given sequence segment of width w matched a motif by a function $S = \max(S_+, S_-)$ where S_+ and S_- are the following scoring formula on the sequence itself and its reverse complement, respectively:

$$S = \sum_{i=1}^{w} \sum_{j \in \{A,C,G,T\}} \delta_{ij} \ln\left(\frac{p_{ij} + p_s}{b_j}\right)$$

where p_{ij} is the frequency of nucleotide j at position i in the motif, p_s is a pseudocount of 0.03, b_j is the background probability of nucleotide j calculated from the intergenic regions of the human genome, $\delta_{ij} = 1$ if nucleotide j is present at position i and $\delta_{ij} = 0$ if nucleotide j is not present at position i. Since the p53 is known to bind two palindrome 10mers separated by a variable spacer of length 0–13 bp (el-Deiry *et al*., 1992), we computed the p53 matches by summing the scores for the two matching 10mers and extended the spacer to 30 bp.

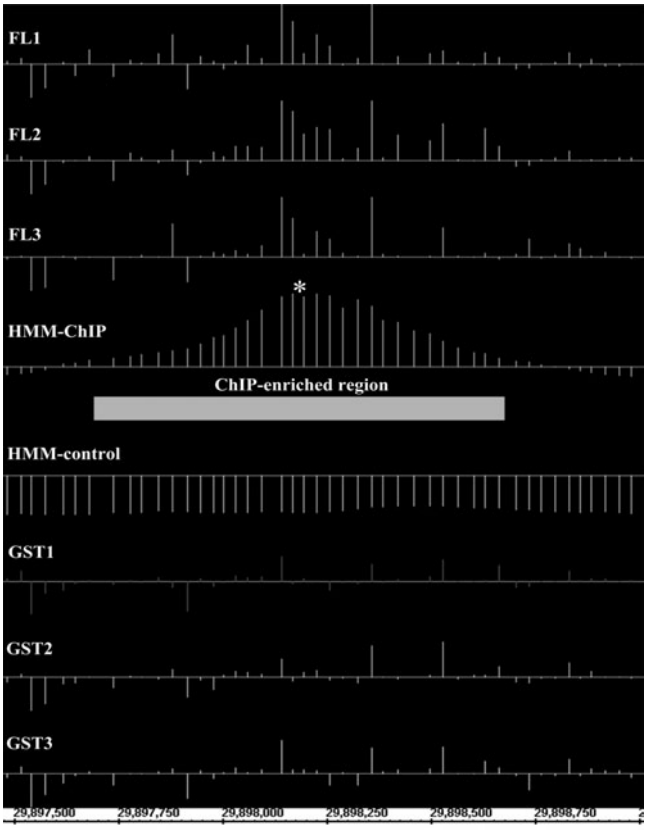

Fig. 3. Typical ChIP-enriched region (Blk78). Quantile normalized (PM−MM) probe values for 3 ChIPs (p53-FL) and 3 controls (GST) as well as HMM log-odds enrichment value for ChIP and control on each probe were mapped to the chromosome 22 positions (bottom line). Although the HMM log-odds enrichment values are from six ChIPs and six controls, only three ChIPs and three controls are shown in this figure. The height of the vertical bar is proportional to either probe signal or HMM enrichment value with the horizontal line indicating the value 0. The ChIP enriched-region with default cutoff was indicated by the rectangle in the middle. HMM fits the data into a smooth, symmetrical, bell-shaped curve. The two 10mers p53 binding motif with 0 bp gap (AGACAGGCTC{0}AGGCATGCCA, indicated as asterisk) is right in the middle region of this curve with the highest HMM enrichment value.

3 RESULTS

3.1 Identification of ChIP-enriched sequences

We applied our analysis strategy (Fig. 1) on the data of p53-FL ChIP-chip experiments on chromosomes 21 and 22 tiling arrays (Cawley *et al*., 2004). It was observed that 3.8% of the data representing short-range (<1 kb) repetitive probe measurements were filtered out, leaving behind 1 014 067 non-redundant probe measurements. With the HMM algorithm described in Section 2.4, we identified 98 p53 TFBSs (defined as HMM-Full TFBSs and summarized in Table 1 in the Supplementary materials), which include all the 10 p53 TFBSs previously verified by quantitative PCR (S. Bekiranov, personal communication: Cawley *et al*., 2004 reported 11 quantitative PCR verified regions for p53-FL, but actually only 10 regions were verified). Overall, 24 of 98 HMM-Full TFBSs overlap with the 48 p53-FL TFBSs reported by Cawley *et al*. (2004). Furthermore, although we did not use the p53-DO1 data in our HMM analysis, six additional TFBSs in our HMM-Full results agree with the p53-DO1 result from Cawley *et al*. (2004). In a typical ChIP-enriched region (Fig. 3), HMM will fit the data into a smooth, symmetrical, bell-shaped curve. Often, the p53 binding motif appears in the middle region of the curve with the highest HMM value.

There are two major groups of repeats in eukaryotic genomes: interspersed repeats mainly represent degenerate

copies of transposable elements dispersed at various locations, whereas tandem repeats are usually confined to specific genomic regions where a unit is tandemly repeated almost exactly from several to thousands of times. To minimize potential cross-hybridization, current Affymetrix probe design (Kapranov *et al*., 2002) rejects probes residing in the interspersed repeats and tandem repeats with short period (roughly $\leq$12) identified by RepeatMasker. Even though we filter out the repetitive probe measurements, each repeat unit can still be represented by probe(s) with different 25mers. Therefore, regions ($\sim$1% of the genome) containing tandem repeats with long period ($>$12) and high copy number ($>$10) will have more chance to be falsely predicted as ChIP-enriched. This false-positive prediction includes both higher probe signal value than the real one and expanding real enriched area to the entire tandem repeat region. In one example, Blk55 (Table 1 in the Supplementary materials), containing a tandem repeat with period size of 54 bp and copy number of 139, is $\sim$7 kb long and has extremely high enrichment score of 14.7. This region, comprising $\sim$160 probes, was not repeat-masked during the array design and might be falsely predicted as ChIP-enriched. We identified a total of four TFBSs within the tandem repeats (Table 1 in the Supplementary materials). Among the 21 TFBSs from Cawley *et al*. (2004) that were not identified by our HMM approach, almost half are within the tandem repeats. They may indicate a higher number of false positive predictions by Cawley *et al*. (2004), although they could be involved in various regulatory mechanisms (Nakamura *et al*., 1998). We decided to keep all the HMM-identified TFBSs within the tandem repeats, but only use the fully-repeat-masked sequences for downstream *de novo* motif finding.

In addition to the known repeats, non-RepeatMasked large segmental duplications ($>$90% identity, $>$1 kb in length) cover $\sim$5.3% of the euchromatic genome from the current human genome assembly (International Human Genome Sequencing Consortium, 2004). This is the most difficult cross-hybridization problem for ChIP-chip experiments on genome tiling arrays. Without other independent evidence, it is impossible to discriminate one or more copies of the real enriched DNA from the large segment duplications based solely on tiling array hybridization. For example, one $\sim$40 kb segment duplicates six times within chromosome 22 with $\sim$99% sequence identity. We found two TFBSs (the first occurrences are Blk35 and Blk36) on each copy of this duplicated segments, generating 12 TFBSs 'redundant' at sequence identity level. A total of 26 from the 98 HMM-Full TFBSs were found within these segmental duplications and were kept intact in our current analysis.

3.2 Identification of p53-binding motif

Extracting putative regulatory motifs from ChIP-enriched regions is difficult for genome tiling array data because the long enriched sequences increase the background noise.

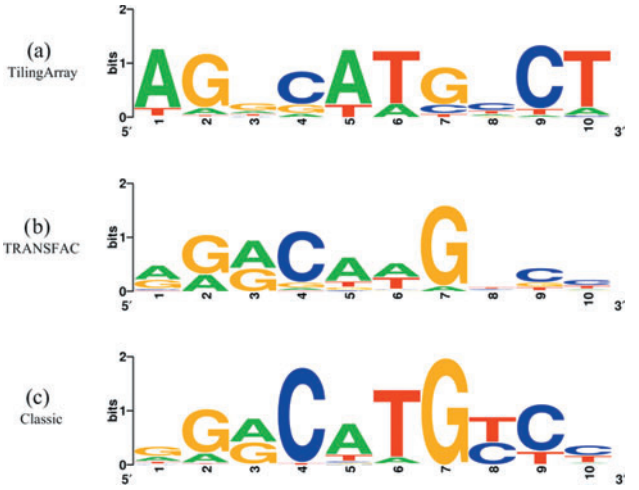

Fig. 4. 10mer sequence logos of p53 motifs from (**a**) MDscan identified motif from 43 HMM-identified high confidence regions from p53 ChIP-chip experiments on genome tiling arrays; (**b**) TRANSFAC Professional database; (**c**) aligned p53 binding sequences from two Classic literatures (el-Deiry *et al*., 1992; Funk *et al*., 1992) defining the p53 consensus binding site.

Therefore, we carried out motif finding only on the 43 high confidence regions with a stringent log-odds enrichment cutoff value of 6 in ChIP and the same default cutoff in control and defined these TFBSs containing high confidence regions as HMM High-Confidence (Table 1 in the Supplementary materials). Using a motif-finding program MDscan (Liu *et al*., 2002), we successfully recovered a strong 10mer palindromic binding motif (Fig. 4 TilingArray motif) from these 43 fully-repeat-masked high-confidence regions, This motif resembles both the p53 motif from entry M00761 of TRANSFAC (Matys *et al*., 2003) and the Classic p53 motif derived by aligning p53 binding sequences from two literatures (el-Deiry *et al*., 1992; Funk *et al*., 1992). A similar TilingArray motif could still be recovered after the segmental duplication sequences were removed from the high-confidence regions. It is the first time a p53 motif is successfully predicted from either promoters of co-expressed genes or ChIP-chip enriched sequences by *de novo* motif finding algorithms. Biologists often use the prior knowledge of the TFBS motif from literature or databases to search for the occurrences of the binding sites. However, if the motif is obtained only from several known sites, it may be either too restrictive and miss many real binding sites or too general and find many false positive sites. In contrast, the p53 TilingArray motif identified in this study might represent the unbiased characterization of p53-DNA interaction. For example, the Classic p53 binding motif require a C at position 4 and a G at position 7, whereas the TilingArray motif is somewhat degenerate at these two positions. Some studies (Resnick-Silverman *et al*., 1998; Jaiswal and Narayan, 2001) showed p53 to bind to sequences with mutations in these two positions, indicating

Table 1. Enrichment of p53 binding motifs

Motif	TFBS	No. of sites	No. of bases	No. of motifs	% Sites with motif	No. of motifs per 10 kb	Fold-enrichment	Binomial *P*-value
TilingArray	HMM-High-Confidence	43	31,394	34	40	10.8	3.3	2.3e−09
	HMM-Full	98	58,734	55	34	9.4	2.8	4.8e−12
	Cawley's	48	25,994	20	25	7.7	2.3	2.1e−04
TRANSFAC	HMM-High-Confidence	43	31,394	73	60	23.3	3.5	<2.2e−16
	HMM-Full	98	58,734	101	42	17.2	2.6	<2.2e−16
	Cawley's	48	25,994	35	31	13.5	2.0	3.0e−05
Classic	HMM-High-Confidence	43	31,394	26	35	8.3	4.6	2.8e−10
	HMM-Full	98	58,734	36	23	6.1	3.4	3.0e−10
	Cawley's	48	25,994	13	17	5.0	2.8	6.3e−4

Three p53 motifs (Fig. 4) were mapped to three sets of p53 TFBSs (Table 1 in the Supplementary materials). The number of bases refers to the number of non-repeat nucleotides in fully-repeat-masked TFBSs. A motif-matching scoring function with cutoff of 10, allowing up to 30 bp spacer between the 10mer motif pairs, was used to determine the number of matches in individual TFBS. This cutoff corresponds to 3.3 matches per 10 kb on chromosomes 21 and 22. Fold enrichment was inferred by comparing the motif occurrences in TFBSs with those in fully-repeat-masked chromosomes 21 and 22 sequences. A one tail binomial test was used to determine the *P*-value attached to the motif enrichment. A substantial enrichment of p53 motif was observed in our HMM identified TFBS.

that the TilingArray p53 motif might better characterize the p53 binding at these two positions than the Classic p53 motif.

3.3 Motif enrichment

p53 was known to bind to two copies of the 10mer palindrome motif with a variable spacer between them (el-Deiry *et al*., 1992). We mapped the p53 motif occurrences to the fully-repeat-masked sequences of chromosomes 21 and 22, allowing up to 30 bp spacer between the 10mer motif pairs. A substantial enrichment of the TilingArray motif pairs was observed from our HMM identified TFBSs. Using a match score cutoff of 10 (corresponding to 3.3 matches per 10 kb on chromosomes 21 and 22), TilingArray motif pairs are in 40, 34 and 25% of HMM-High-Confidence, HMM-Full and Cawley's TFBS, respectively, corresponding to 3.3-, 2.8- and 2.3-fold enrichment compared with the genomic background (Table 1). In addition, in the TFBSs lacking TilingArray motif pairs, 10 HMM-High-Confidence and 21 HMM-Full were found to contain at least one copy of the TilingArray motif. The single copy TilingArray motif (not motif pairs) represents a 2.7-fold enrichment compared with the genomic background. One might suspect the validity of this enrichment analysis, since the TilingArray motif itself is extracted from the HMM identified high confidence regions and thus, should be much more enriched. To address this issue, we conducted enrichment analysis using two independent representations of p53 motif (TRANSFAC and Classic, Fig. 4) with the same criteria on the three sets of TFBSs (Table 1). The result showed that the fold enrichments of these two p53 motifs in our HMM identified regions are still much higher than those in Cawley's data.

3.4 Annotation of p53 TFBS

We examined the proximities of p53 TFBSs to traditional transcriptional regulatory regions, such as CpG islands and 5′ upstream or 3′ downstream of RefSeq genes (Table 2). A total of 5, 6 and 8 TFBSs in HMM-High-Confidence, HMM-Full and Cawley's, respectively, are within 1 kb of annotated CpG island. This seems to indicate that more Cawley's TFBSs are proximal to CpG island than those identified by HMM. However, since TFBSs <1 kb in Cawley *et al*. (2004) were extended equally in both directions to a length of 1 kb, they are more likely to be longer than ours. When the distance between TFBS and CpG island was expanded to 3 kb, we found almost the same number of TFBS in HMM-High-Confidence and in Cawley's TFBSs. It is worth noting that no TFBS of Cawley are <1 kb of 5′ upstream or 3′ downstream of RefSeq gene, whereas three TFBSs identified in HMM are within those regions. The one TFBS found to be within 1 kb of 3′ downstream of RefSeq gene is of special interest, because it suggests the potential antisense transcripts at the 3′ of gene.

We extracted 40 well-documented p53 directly regulated genes from TRANSFAC (Matys *et al*., 2003) and mapped them to the human genome. Unfortunately, none of them is located along human chromosomes 21 and 22. Recently,

Table 2. Annotation of p53 TFBS locations

TFBS	No. of sites	CpG island <1 kb	CpG island <3 kb	RefSeq gene <1 kb to 5′	RefSeq gene <1 kb to 3′
HMM-High-Confidence	43	5	10	1	1
HMM-Full	98	6	20	2	1
Cawley's	48	8	12	0	0

The three sets of TFBS sets are the same as in Table 1 in the Supplementary materials. CpG island and RefSeq Gene annotation tracks were retrieved from UCSC Genome Browser on April 2003 Human Genome Assembly (http://genome.ucsc.edu/). 5′ and 3′ referred to the transcription starts and ends in RefSeq Gene track. The distance between TFBS and CpG island or RefSeq Gene is the space between their nearest boundaries.

two studies (Mirza *et al*., 2003; Kho *et al*., 2004) identified thousands of p53-target genes (direct or indirect) from microarray expression analysis. They further narrowed down the potential p53 directly regulated genes to several hundred by finding p53-bindng consensus sequence in their regulatory regions (Mirza *et al*., 2003) or by excluding the genes whose expression are not directly influenced by p53 protein level (Kho *et al*., 2004). Thirty of these potential p53 directly regulated genes are along chromosomes 21 and 22. We found two TFBSs to be associated with these genes: Blk73, which was not identified by Cawley *et al*. (2004), is in the intron of known gene *PITPNB* (protein gi|1060905) with a 2.8-expression-fold change (Mirza *et al*., 2003) in response to p53; Blk87, which was verified by quantitative PCR in Cawley *et al*. (2004), is in the ∼13 kb upstream of known gene *AB051455* (protein gi|6572156) with −4.1-expression-fold change (Mirza *et al*., 2003) in response to p53.

3.5 Method evaluation

Our analysis strategy (Fig. 1) differs from the previous study by Cawley *et al*. (2004) in data normalization, repeat elements filtering and HMM prediction. It is worth checking whether the better performance of binding sequence prediction is due to the HMM prediction or the other two factors. To investigate the real difference between HMM and Mann–Whitney U-test, we performed the Mann–Whitney U-test on our quantile normalized datasets with repetitive probe measurements removed. We used the same criteria from Cawley *et al*. (2004), i.e. ±500 bp sliding window; probe values were transformed to $\log 2(\max(PM-MM,1))$; predicted regions separated by <500 bp were merged together; two sets of prediction (FL–DO1, FL–Input) were further merged together to form a non-redundant set. Using the same P-value cutoff of 1e−5, we identified 65 TFBSs, from which additional 11 TFBSs could be found in HMM-Full besides the previous 24 overlapping TFBSs between HMM-Full and Cawley's data. Most of the 11 additional TFBSs fall just barely below the 1e–5 cutoff. We gradually decreased the P-value cutoff from 1e−5 to 1e−9 to conduct the *de novo* p53-binding motif searching on the fully-repeat-masked resulting sequences. With none of these cutoffs could MDscan recover a pattern resembling the p53 consensus binding site. As for the motif enrichment study, we used the 1e−5 P-value cutoff and expanded the resulting TFBS equally in each direction to have a minimum length of 750 bp. Therefore, the total number of non-repeat nucleotides in these 65 TFBSs is ∼34 kb, which is comparable with that (∼31 kb) of HMM-High-Confidence. Using three p53 motif representations with the same criteria as in Table 1, we found 7.6, 15.7 and 6.4 motifs per 10 kb for TilingArray, TRANSFAC and Classic p53 motifs, respectively. The motif enrichment is much lower than that in our HMM-High-Confidence regions. All of the above results suggest that HMM contributed greatly to the success of our binding sequence prediction.

4 DISCUSSION

We present a fast, scalable and sensitive HMM approach for analyzing ChIP-chip experiments on genome tiling arrays. The algorithm is fast because replicate datasets at each probe location are considered together in a single HMM run to estimate its enrichment probability. On a 2.4 MHz Xeon server, it takes ∼6 min to run HMM on six ChIP and six control replicates each with ∼1 million probes. Because of its speed, the algorithm is scalable on tiling arrays of either small regions or the whole genome, and flexible enough to be used on other organisms. Furthermore, the algorithm is sensitive in identifying all the regions previously verified by quantitative PCR. The following independent evidences suggest that the 98 TFBSs we identified are likely to be genuine regulatory elements. First, our analysis strategy reported fewer TFBSs residing in non-RepeatMasked tandem repeat and large segmental duplications regions which are questionable in their regulatory function. Second, a TilingArray motif resembling the TRANSFAC and Classic p53 motif was successfully discovered from the high confidence TFBSs. Finally, matching the TilingArray, TRANSFAC and Classic p53 motifs in the TFBSs reveals substantial enrichment of the motifs compared with the genomic background.

Our algorithm overcomes the intrinsic weakness of the Mann–Whitney U-test used in Cawley *et al*. (2004). Without enough replicates and long enough window, Mann–Whitney U-test cannot identify enriched regions with a sufficiently small P-value. On account of the high cost of the tiling arrays, often biologists cannot afford more than three replicates. The shearing process in ChIP usually generates fragments <1 kb long. Reinforcing a 1 kb window may not only miss short ChIP-enriched regions but also include sequences outside the enriched regions which may confound later motif detection. Mann–Whitney U-test treats every probe equally and fails to consider probe-by-probe variability. The statistic U does not reflect whether probe values are fluctuating a lot or continuously high within the window; only the latter indicates ChIP-enriched TFBS.

Our HMM approach is based on normalizing data on tiling arrays from many experiments and modeling the behavior of each individual probe. One laboratory may not necessarily have enough ChIP-chip replicates, but pooling together freely available tiling array data from many different laboratories and different experiments (e.g. ChIP-chip against different TFs or different tissues) can provide enough information for the modeling requirement. Probe behavior estimated from pooled experiments serves as a baseline for each ChIP experiment to be compared with, thus HMM can even get reasonable results from a single ChIP experiment. This is particularly useful for surveying at the beginning of a ChIP experiment to explore antibody quality, culture conditions or preliminary quality assessment of each replicate. Our HMM approach calculates enrichment probability at each probe location, so the

exact enriched region boundaries could be determined. Furthermore, the sensitive HMM approach can identify short ChIP-enriched sequence. If we exclude those TFBS containing tandem repeats, the average length (771 bp) of the HMM-High-Confidence TFBSs not identified by the Mann–Whitney *U*-test is much shorter than the average length (984 bp) of HMM-High-Confidence TFBSs identified by Cawley *et al*. (2004).

Our analysis seems to reveal two interesting characteristics about p53 regulation. First, p53 binding may not simply be the classic two 10mer palindrome motif separated by a variable length spacer of 0–13 bp. Many HMM identified TFBSs contain only a single copy of the TilingArray motif or multiple copies that are >30 bp apart. In almost half (43) of the HMM-Full TFBSs, including four quantitative PCR verified regions, this motif does not occur at all. The absence of known p53 motif or motif pairs in TFBS suggests that there might be an alternative binding mechanism (Yin *et al*., 2003) for p53. Second, p53 binding sites may not be conserved between human and rodents. The average sequence identity for all the 98 HMM-Full TFBSs is only ∼30% based on the BLASTZ global alignment .(Schwartz *et al*., 2003) between human (hg15) and rodents (either mouse mm3 or rat rn2). Even the average sequence identity of all the quantitative PCR verified regions is only 42%. Furthermore, 37 TFBSs, including one quantitative PCR verified region, have no rodent counterparts at all. Many algorithms for finding eukaryotic regulatory elements rely on identifying conserved sequences from alignment of orthologous non-coding sequences (Loots *et al*., 2002; Liu *et al*., 2004). Our result suggests that the loss of sensitivity of this approach may be significant. The majority of the TFBSs identified by ChIP-chip experiments on tiling arrays will be sacrificed if sequence identity is included as a criterion.

In summary, the HMM approach presented here provides biologists with an efficient computational approach for analyzing the massive data generated from ChIP-chip on genome tiling arrays. Its adoption will contribute to the more comprehensive models and understanding of gene regulatory networks in higher eukaryotes.

ACKNOWLEDGEMENTS

The authors would like to thank David Harrington, Chen Li, Molin Wang and Jun S. Liu for their insights and advice on the tiling array data analysis. The project was partly supported by Claudia Adams Barr Program in Cancer Research.

REFERENCES

Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.

Bolstad,B.M., Irizarry,R.A., Astrand,M. and Speed,T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Cawley,S., Bekiranov,S., Ng,H.H., Kapranov,P., Sekinger,E.A., Kampa,D., Piccolboni,A., Sementchenko,V., Cheng,J., Williams,A.J. *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.

Conlon,E.M., Liu,X.S., Lieb,J.D. and Liu,J.S. (2003) Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl Acad. Sci. USA*, **100**, 3339–3344.

el-Deiry,W.S., Kern,S.E., Pietenpol,J.A., Kinzler,K.W. and Vogelstein,B. (1992) Definition of a consensus binding site for p53. *Nat. Genet.*, **1**, 45–49.

Euskirchen,G., Royce,T.E., Bertone,P., Martone,R., Rinn,J.L., Nelson,F.K., Sayward,F., Luscombe,N.M., Miller,P., Gerstein,M., Weissman,S. and Snyder,M. (2004) CREB binds to multiple loci on human chromosome 22. *Mol. Cell. Biol.*, **24**, 3804–3814.

Funk,W.D., Pak,D.T., Karas,R.H., Wright,W.E. and Shay,J.W. (1992) A transcriptionally active DNA-binding site for human p53 protein complexes. *Mol. Cell. Biol.*, **12**, 2866–2871.

Harbison,C.T., Gordon,D.B., Lee,T.I., Rinaldi,N.J., Macisaac,K.D., Danford,T.W. Hannett,N.M., Tagne,J.B., Reynolds,D.B., Yoo,J. *et al*. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.

Jaiswal,A.S. and Narayan,S. (2001) p53-dependent transcriptional regulation of the APC promoter in colon cancer cells treated with DNA alkylating agents. *J. Biol. Chem.*, **276**, 18193–18199.

Kapranov,P., Cawley,S.E., Drenkow,J., Bekiranov,S., Strausberg,R.L., Fodor,S.P. and Gingeras,T.R. (2002) Large-scale transcriptional activity in chromosomes 21 and 22. *Science*, **296**, 916–919.

Kho,P.S., Wang,Z., Zhuang,L., Li,Y., Chew,J.L., Ng,H.H., Liu,E.T. and Yu,Q. (2004) p53-regulated transcriptional program associated with genotoxic stress-induced apoptosis. *J. Biol. Chem.*, **279**, 21183–21192.

Li,Z., Van Calcar,S., Qu,C., Cavenee,W.K., Zhang,M.Q. and Ren,B. (2003) A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. *Proc. Natl Acad. Sci. USA*, **100**, 8164–8169.

Lieb,J.D., Liu,X., Botstein,D. and Brown,P.O. (2001) Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat. Genet.*, **28**, 327–334.

Lieberfarb,M.E., Lin,M., Lechpammer,M., Li,C., Tanenbaum,D.M., Febbo,P.G., Wright,R.L., Shim,J., Kantoff,P.W., Loda,M., Meyerson,M. and Sellers,W.R. (2003) Genome-wide loss of heterozygosity analysis from laser capture microdissected prostate cancer using single nucleotide polymorphic allele (SNP) arrays and a novel bioinformatics platform dChipSNP. *Cancer Res.*, **63**, 4781–4785.

Liu,X.S., Brutlag,D.L. and Liu,J.S. (2002) An algorithm for finding protein–DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.*, **20**, 835–839.

Liu,Y., Liu,X.S., Wei,L., Altman,R.B. and Batzoglou,S. (2004) Eukaryotic regulatory element conservation analysis and identification using comparative genomics. *Genome Res.*, **14**, 451–458.

Loots,G.G., Ovcharenko,I., Pachter,L., Dubchak,I. and Rubin,E.M. (2002) rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.*, **12**, 832–839.

Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al*. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.

Mirza,A., Wu,Q., Wang,L., McClanahan,T., Bishop,W.R., Gheyas,F., Ding,W., Hutchins,B., Hockenberry,T., Kirschmeier,P., Greene,J.R. and Liu.S. (2003) Global transcriptional program of p53 target genes during the process of apoptosis and cell cycle progression. *Oncogene*, **22**, 3645–3654.

Nakamura,Y., Koyama,K. and Matsushima,M. (1998) VNTR (variable number of tandem repeat) sequences as transcriptional, translational, or functional regulators. *J. Hum. Genet.*, **43**, 149–152.

Odom,D.T., Zizlsperger,N., Gordon,D.B., Bell,G.W., Rinaldi,N.J., Murray,H.L. Volkert,T.L., Schreiber,J., Rolfe,P.A., Gifford,D.K. *et al*. (2004) Control of pancreas and liver gene expression by HNF transcription factors. *Science*, **303**, 1378–1381.

Rabiner,L. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–286.

Ren,B., Robert,F., Wyrick,J.J., Aparicio,O., Jennings,E.G., Simon,I., Zeitlinger,J., Schreiber,J., Hannett,N., Kanin,E. *et al*. (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.

Resnick-Silverman,L., St Clair,S., Maurer,M., Zhao,K. and Manfredi,J.J. (1998) Identification of a novel class of genomic DNA-binding sites suggests a mechanism for selectivity in target gene activation by the tumor suppressor protein p53. *Genes Dev.*, **12**, 2102–2107.

Schwartz,S., Kent,W.J., Smit,A., Zhang,Z., Baertsch,R., Hardison,R.C., Haussler, D. and Miller, W. (2003) Human–mouse alignments with BLASTZ. *Genome Res.*, **13**, 103–107.

Wells,J., Yan,P.S., Cechvala,M., Huang,T. and Farnham,P.J. (2003) Identification of novel pRb binding sites using CpG microarrays suggests that E2F recruits pRb to specific genomic sites during S phase. *Oncogene*, **22**, 1445–1460.

Workman,C., Jensen,L.J., Jarmer,H., Berka,R., Gautier,L., Nielser,H.B., Saxild,H.H., Nielsen,C., Brunak,S. and Knudsen,S. (2002) A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol.*, **3**, research0048.

Yin,Y., Liu,Y.X., Jin,Y.J., Hall,E.J. and Barrett,J.C. (2003) PAC1 phosphatase is a transcription target of p53 in signalling apoptosis and growth suppression. *Nature*, **422**, 527–531.