BIOINFORMATICS ORIGINAL PAPER

Vol. 25 no. 18 2009, pages 2369–2375
doi:10.1093/bioinformatics/btp425

*Gene expression*

# TileProbe: modeling tiling array probe effects using publicly available data

Jennifer Toolan Judy[1] and Hongkai Ji[2,*]

[1]Department of Mental Health and [2]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, MD 21205, USA

Received on April 15, 2009; revised on June 16, 2009; accepted on July 7, 2009

Advance Access publication July 9, 2009

Associate Editor: John Quackenbush

**ABSTRACT**

**Motivation:** Individual probes on an Affymetrix tiling array usually behave differently. Modeling and removing these probe effects are critical for detecting signals from the array data. Current data processing techniques either require control samples or use probe sequences to model probe-specific variability, such as with MAT. Although the MAT approach can be applied without control samples, residual probe effects continue to distort the true biological signals.

**Results:** We propose TileProbe, a new technique that builds upon the MAT algorithm by incorporating publicly available data sets to remove tiling array probe effects. By using a large number of these readily available arrays, TileProbe robustly models the residual probe effects that MAT model cannot explain. When applied to analyzing ChIP-chip data, TileProbe performs consistently better than MAT across a variety of analytical conditions. This shows that TileProbe resolves the issue of probe-specific effects more completely.

**Availability:** http://www.biostat.jhsph.edu/~hji/cisgenome/index_files/tileprobe.htm

**Contact:** hji@jhsph.edu

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 INTRODUCTION

High density tiling arrays are widely used to study transcription factor binding (Cawley *et al*., 2004; Carroll *et al*., 2005), transcriptome (Bertone *et al*., 2004; Kapranov *et al*., 2002), DNA methylation (Weber *et al*., 2007; Zhang *et al*., 2006), chromatin modification (Bernstein *et al.*, 2006), nucleosome positioning (Yuan *et al.*, 2005; Ozsolak *et al*., 2007) and copy number variations (Urban *et al*., 2006). Among the various array platforms, Affymetrix tiling arrays offer the lowest price per probe, the highest resolution, and can be used in most of the applications above (see Liu, 2007 for a review). These arrays use densely spaced probes to interrogate either the entire or part of the genome. Similar to other microarray platforms, different probes on an Affymetrix tiling array usually behave differently. These probe-specific behaviors, also known as probe effects (Irizarry *et al*., 2003; Johnson *et al*., 2006; Li and Wong, 2001; Wu *et al*., 2004), need to be properly controlled before meaningful biological signals can be extracted from the data. Figure 1a provides an example that illustrates the probe effects in a typical ChIP-chip experiment, in which DNA fragments bound by a transcription factor are collected through chromatin immunoprecipitation (ChIP) and hybridized to tiling arrays. The first three tracks show $\log_2$ transformed probe intensities of three independent ChIP samples for a transcription factor Gli3, representing three biological replicates. The next three tracks show $\log_2$ transformed probe intensities of three control samples in which the immunoprecipitation step was skipped. Track 7 shows $\log_2$ fold changes between the ChIP and control intensities averaged across three replicates. The peak in this track is a functional Gli3 binding site that has been experimentally verified. Existence of probe effects is clearly demonstrated by the fact that many probes outside the binding region have higher intensity values than probes inside the binding region (e.g. compare probes highlighted by the boxes), and this trend is consistent across all the samples. A direct consequence of probe effects is that the first three tracks alone (ChIP samples without controls) incorrectly define the location of transcription factor binding.

To handle the probe effects, one class of methods is to compare two groups of samples to yield relative measures. In principle, this is similar to the log fold change track in Figure 1a. These methods first normalize probe intensities across samples and then look for regions where probe intensities are significantly different between two different groups (e.g. ChIP versus control in a ChIP-chip experiment). Some examples in this class include Affymetrix's TAS (Kampa *et al*., 2004), TileMap (Ji and Wong, 2005), HMMTiling (Li *et al*., 2005), and Keles *et al*. (2006). These methods use quantile normalization (Bolstad *et al*., 2003) to match intensities across samples, after which techniques including Wilcoxon rank-sum test, *t* or *t*-like statistics, and hidden Markov model, etc. are used to measure the ChIP/control differences and detect protein–DNA binding regions.

Another popular method, represented by MAT (Johnson *et al*., 2006), uses probe sequences to model the probe effects. The MAT method assumes that most probes on an array measures background noise, an assumption that usually holds in applications such as ChIP-chip. With this assumption, MAT attempts to explain background probe intensities in a single Affymetrix array by fitting a regression, using log probe intensities as the response and probe sequences and probe copy numbers in the genome as the covariates. The difference between the observed and model-fitted log probe intensities, after further processing, forms the background corrected probe intensity. Using the background corrected probe

*To whom correspondence should be addressed.

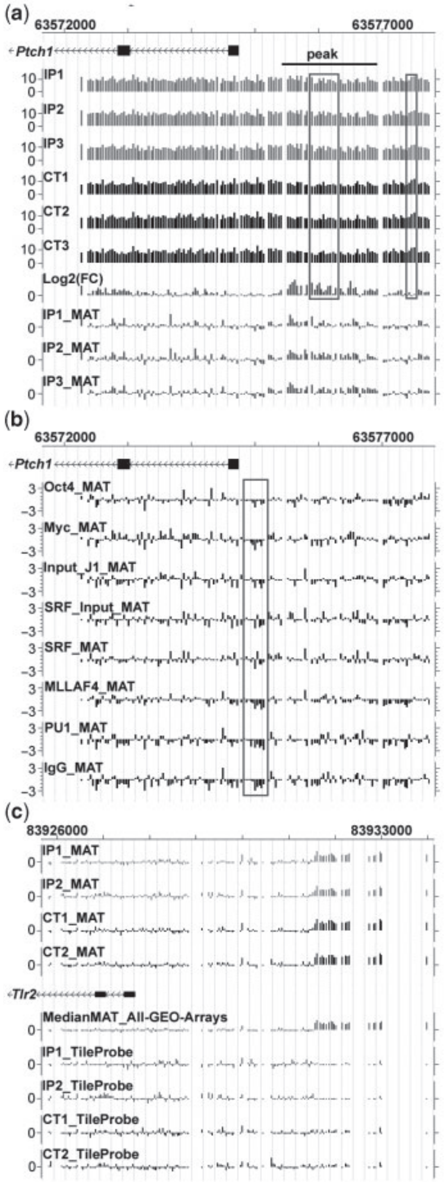

**Fig. 1.** Illustration of probe effects on Affymetrix Mouse Promoter 1.0R arrays. (**a**) IP1–IP3, CT1–CT3: quantile normalized Gli3 ChIP and control probe intensities at $\log_2$ scale. $\text{Log}_2$(FC): $\log_2$(IP/CT) fold change. IP1_MAT-IP3_MAT: MAT background corrected probe intensities for IP1–IP3. (**b**) MAT corrected probe intensities for samples collected from different studies. (**c**) IP_MAT, CT_MAT: MAT corrected probe intensities. MedianMAT_All-GEO-Arrays: median MAT corrected probe intensities across all samples stored in GEO. IP_TileProbe, CT_TileProbe: TileProbe background corrected probe intensities.

intensities allows one to eliminate a significant portion of sequence-dependent probe behaviors, including the increased mean and variance commonly observed in GC-rich probes. As a result, peak signals can be detected even from a single array. This is illustrated by the last three tracks of Figure 1a, in which MAT background corrected probe intensities for the three ChIP samples are displayed. Since MAT can analyze experiments without control samples, it is a very attractive tool in pilot studies (e.g. a study to test antibodies), and in studies limited by cost constraints (e.g. a study that involves profiling a large number of samples and/or conditions).

Although the MAT model allows one to remove a significant portion of probe effects from tiling array raw data, it is unable to remove all probe effects. To illustrate this, we applied MAT to analyze all samples in the GEO database (Barrett *et al*., 2007) collected using Affymetrix Mouse Promoter 1.0R arrays. Figure 1b shows that the MAT background corrected probe intensities have residual probe effects, and the residual probe effects are consistent across different samples collected for different transcription factors by different labs. These residual probe effects could be explained by several factors. First, MAT uses an unsaturated model that includes only the main effects of probe sequence and a few squared terms as covariates. As a result, it cannot explain probe effects due to higher order interactions between nucleotides at different positions within a probe. Second, not all probe effects are sequence dependent (e.g. the physical locations of probe in the array may also contribute to the probe effects). Therefore, the prediction of probe effects based on probe sequences may not be perfect.

The residual probe effects in the MAT corrected probe intensities could directly affect the subsequent detection of biological signals. For example, some probes consistently show negative MAT values across samples (e.g. probes highlighted by the box in Fig. 1b). If there were true biological signals in these regions, but they were not strong enough to reverse the sign of MAT corrected probe intensities, and no control samples were available, MAT would not be able to detect these signals. Figure 1c shows another extreme, i.e. some probes consistently have high background corrected probe intensities. If one only had the ChIP samples in the first two tracks, the region interrogated by these probes would be declared by MAT as a high confidence peak. However, the comparison with control samples clearly illustrates that this region is a false positive. These examples suggest that by further removing the residual probe effects, one should be able to improve the sensitivity and specificity of subsequent analysis.

The observation that MAT background corrected probe intensities have consistent residual probe effects across different samples motivated us to develop a new approach, TileProbe, to handle probe effects of Affymetrix tiling arrays. TileProbe takes advantage of the diverse and large number of samples stored in the GEO database and uses these publicly available data to obtain a robust model for MAT residual probe effects. This method allows one to remove the residual probe effects from the MAT background corrected probe intensities (Fig. 1c). As a result, the issues illustrated in Figure 1b and c can be resolved even without control samples. TileProbe was tested using a number of different ChIP-chip data sets. The tests showed that with the improved probe effects model, TileProbe performed better than MAT across a variety of analytical conditions, including analyses with and without control samples. In some scenarios, the improvement was dramatic.

## 2 METHODS

TileProbe consists of two parts: building the probe model and applying the model to analyze new data. For a given array platform, the first part involves building a probe effect model for each probe by collecting and analyzing existing samples from the GEO database. The second part involves the application of the model to new data sets. The output is background corrected probe intensities, which can be used as input for various subsequent analyses such as peak detection.

## 2.1 Building probe model

To build the model, we first apply MAT to each individual sample collected from GEO. In other words, for each sample, the following regression is fitted using all perfect match probes on the array:

$$\mathrm{Log}(PM_i)=\alpha n_{iT}+\sum_{j=1}^{25}\sum_{k\in\{A,C,G\}}\beta_{jk}I_{ijk}+\sum_{k\in\{A,C,G,T\}}\gamma_k n_{ik}^2+\delta\log(c_i)+\varepsilon_i \tag{1}$$

Each probe on the array is 25-bp long. Consistent with the original MAT (Johnson *et al.*, 2006), $PM_i$ is the intensity of probe $i$; $n_{ik}$ is the number of nucleotide $k$ in probe $i$; $I_{ijk}$ indicates whether the $j$th nucleotide of probe $i$ is $k$ ($I_{ijk}$=1) or not ($I_{ijk}$ = 0); $c_i$ is the number of times the sequence of probe $i$ occurs in the genome; $\alpha, \beta_{jk}$, $\gamma_k$ and $\delta$ are regression coefficients; and $\varepsilon_i$ is the probe specific error. Using the fitted parameters, log probe intensity of probe $i$ can be predicted as $\hat{m}_i$. Probes with similar $\hat{m}_i$ are grouped into affinity bins, each containing 3000 probes. Let $s_i$ be the standard deviation of the affinity bin containing probe $i$. The MAT corrected probe intensity is

$$t_i=\frac{\log(PM_i)-\hat{m}_i}{s_i} \tag{2}$$

The statistic $t_i$ removes a major fraction of sequence dependent probe intensities and probe variances. After this step, a MAT corrected intensity $t_i$ is attached to each probe for a particular sample.

In the second step of TileProbe model building, all samples from a given array platform in GEO are grouped according to studies and experimental conditions. For example, if a study (determined by the GEO series number) contains three ChIP samples and three control samples, the six samples will be divided into two groups: an IP group and a control group. Assume that there are $G$ groups in total and group $g$ ($g\in\{1,2,...,G\}$) contains $K_g$ replicate samples. Let $t_{igk}$ denote the MAT corrected probe intensity of probe $i$ in the $k$-th replicate of group $g$, $\bar{t}_{ig}=\sum_k t_{igk}/K_g$, and $v=\sum_g(K_g-1)$. TileProbe models the residual probe effects in $t_{igk}$ using two quantities $\theta_i$ and $\tau_i$ which are determined as follows:

$$\theta_i=\text{ median }\{t_{igk}\text{ for all }g\in\{1,2,...,G\}\text{ and }k\in\{1,...,K_g\}\}, \tag{3}$$

$$\omega_i^2=\sum_{g=1}^{G}\sum_{k=1}^{Kg}\frac{(t_{igk}-\bar{t}_{ig})^2}{v}, \tag{4}$$

$$\tau_i^2=(1-B)\omega_i^2+B\overline{\omega^2}. \tag{5}$$

Here $\overline{\omega^2}$ is the mean of all $\omega_i^2$, and $B\in[0,1]$ is a shrinkage factor determined by the variance shrinkage estimator used in TileMap (Ji and Wong, 2005):

$$B=\min\left[1,\frac{2}{v+2}\frac{N-1}{N}+\frac{2}{v+2}\frac{(N-1)\left(\overline{\omega^2}\right)^2}{\sum_{i=1}^{N}\left(\omega_i^2-\overline{\omega^2}\right)^2}\right] \tag{6}$$

In the formula above, $N$ is the total number of probes. If $\omega_i^2$ is assumed to follow a chi-square distribution $\omega_i^2|\sigma_i^2\sim\ \sigma_i^2\chi_v^2/v$, and $\sigma_i^2$ values are independent samples from an inverse chi-square distribution $\sigma_i^2\sim$ Inv-$\chi^2(v_0,\ \omega_0^2)$, then the shrinkage factor $B$ represents an estimate of E[Var($\omega_i^2|\sigma_i^2$)]/Var[$\omega_i^2$], and $\tau_i^2$ represents an estimate of probe specific variance $\sigma_i^2$.

To summarize, TileProbe uses $\theta_i$, the median MAT corrected probe intensity across all samples, to model the magnitude of each residual probe effect. This approach requires the assumption that, at each probe, most samples used for building the probe model do not contain biologically relevant signals. This assumption usually holds true when a large number of diverse samples, representing different experimental systems (e.g. different transcription factors in ChIP-chip experiments) and different conditions, are used for model building. In addition, it uses $\tau_i$ to model the probe specific variability expected in a single experimental condition. The shrinkage estimator (5) avoids unstable variance estimates when the available degrees of freedom $v=\sum_g(K_g-1)$ are small.

## 2.2 Applying the model to new data

For each array platform, a probe effect model can be built according to formula (1)–(6). To analyze a new data set generated by the same platform, we first apply MAT correction [i.e. formula (1)–(2)] to each sample, $u$. Next, the MAT corrected probe intensity, $t_{iu}$ for probe $i$ and sample $u$, is standardized as follows:

$$y_{iu}=\frac{t_{iu}-\theta_i}{\tau_i}. \tag{7}$$

The $y_{iu}$ statistic is the TileProbe background corrected probe intensity, which can be used as input for various subsequent analyses such as peak detection.

For test purpose, we also developed several variants of TileProbe. In one variant, formula (7) is replaced by $y_{iu}=t_{iu}-\theta_i$ to test the role of $\tau_i$. This simplified version of TileProbe will be denoted as TPM below, and the original version with variance standardization [i.e. formula (7)] is denoted as TPV. In another variant, quantile normalized log probe intensities replace the MAT corrected probe intensities $t_i$ obtained from formula (1) and (2), after which formula (3)–(7) are applied. This variant is called TPQ and was used to test the role of the MAT correction.

# 3 IMPLEMENTATION

TileProbe is implemented using ANSI C and is incorporated as part of CisGenome (http://www.biostat.jhsph.edu/~hji/cisgenome) (Ji *et al.*, 2008). Precompiled probe models are provided for several commonly used Affymetrix tiling array platforms provided that GEO has enough number of samples for that platform. As a result, users often do not need to run the model-building step themselves, although experienced users can choose to build their own models whenever new samples become available.

# 4 RESULTS

We tested TileProbe using four different ChIP-chip data sets collected from GEO, representing four different transcription factors (TF)—Gli3, Myc, estrogen receptor (ER) and NRSF—and two different array platforms (Affymetrix Mouse Promoter 1.0R and Affymetrix Human Tiling 2.0R array 6) (Table S1). Gli3, Myc and ER each had three ChIP (IP) and three control (CT) samples available, and NRSF had two IP and two CT samples. We examined each TF under the following analytical conditions: 1IP 0CT, 1IP 1CT, 3IP 0CT (2IP 0CT for NRSF), and 3IP 3CT (2IP 2CT for NRSF). The single sample conditions results (i.e. 1IP 0CT, 1IP 1CT) were reported as averages from all available arrays.

Before testing, probe effect models were built using GEO samples that excluded the testing data. TileProbe was then applied to each transcription factor. Using the background corrected probe intensities, TF binding regions were identified using the peak detection procedure described in MAT. Briefly, for each probe, a 600 bp flanking window was formed, and a MATscore was computed as $\sqrt{n}\times TM(y)$, where $n$ was the number of data points in the window, and $TM(y)$ was the trimmed mean of the background corrected probe intensities within the window. The trimmed mean removed the top 10% and bottom 10% of $y$ values, and windows with less than ten probes were excluded from the analysis. If control samples were available, $TM(y)$ was replaced by the difference between the trimmed mean of IP and CT samples. Probes with a MATscore bigger than a user specified cutoff were used to predict binding regions. The peak detection results based on TileProbe corrected intensities were then compared with the results based on MAT corrected intensities

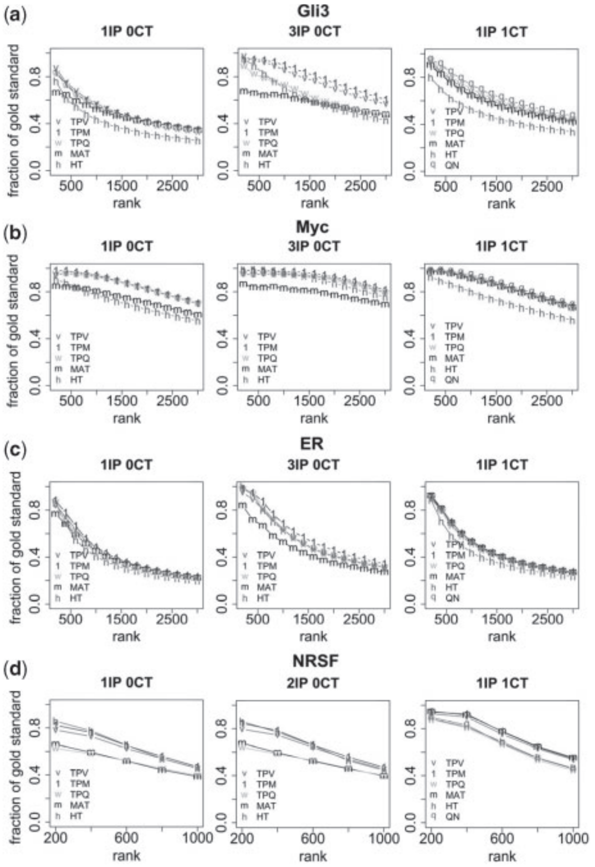

**Fig. 2.** Consistency test. TileProbe (TPV), two variants of TileProbe (TPM and TPQ), MAT and HMMTiling (HT) are compared. For 1IP 1CT, results based on quantile normalization (QN) are also shown. The fraction of predictions that are gold standard is shown for top 200, 400, 600, ... , etc. peaks. The gold standard was constructed using MAT 3IP 3CT analysis. To avoid bias caused by peak length, all peaks were forced to be 500 bp long around the peak maxima.

(i.e. results based on $TM(y)$ and $TM(t)$ were compared). Since we used the same peak detection protocol and the same BPMAP files (downloaded from http://liulab.dfci.harvard.edu/MAT/) for all analyses, the differences observed in the performance comparison are purely due to differences in background correction procedures.

## 4.1 Consistency test

We first compared TileProbe (TPV) and MAT. In the first test, we selected MAT peaks from the fullest available analysis (3IP 3CT or 2IP 2CT) that had a false discovery rate $\leq$10%. We labeled these peaks as the ‘gold standard’ with the highest likelihood of being true TF binding sites. If the gold-standard list contained more than 3000 peaks, only the top 3000 were kept. Next, we examined the number of peaks from the other analytical conditions that overlapped with these gold-standard regions. Figure 2 shows the fraction of gold-standard regions detected by different algorithms among the top-ranked peaks.

For analytical conditions with no control samples (1IP 0CT and 3IP 0CT), TileProbe (TPV) clearly outperformed MAT. Interestingly, the improvement of TPV was stronger when more IP samples were available. This was because with more IP samples, MAT assigned higher confidence level to peaks similar to Figure 1c which were indeed false positives. In some scenarios, the improvement was dramatic. For example, in the 3IP 0CT Gli3 analysis, only ~60% of top 500 MAT predictions overlapped with MAT gold standard, whereas ~90% of top 500 TPV predictions overlapped with MAT gold standard.

In the presence of control samples (1IP 1CT), TPV and MAT performed similarly. However, it should be pointed out that the gold standard was constructed in favor of MAT, and may potentially obscure the improvements of TileProbe.

The MAT used by TileProbe was a reimplementation of Johnson *et al.* (2006). To exclude the possibility that TileProbe’s observed advantage was due to differences in implementing the MAT model and peak detection algorithm, the MAT curves shown in Figure 2 were based on peak detection results obtained using the MAT we implemented in TileProbe. We also compared our implementation of MAT with the original MAT implemented by Johnson *et al.* (2006), and the two versions of MAT performed essentially the same, leaving the conclusions unchanged (Fig. S1).

## 4.2 Motif enrichment test

In the second test, we evaluated the transcription factor binding motif enrichment in the predicted TF binding regions. Since motifs represent an independent source of information, this test does not a priori favor any algorithms over the others. All four TFs have well known motifs that were reported previously and that were constructed using independent data (Table S2 and Fig. S2). Using CisGenome (Ji *et al.*, 2008), the motifs were mapped to the predicted peak lists from different algorithms as well as matched negative control regions. To avoid potential bias due to the peak length, we restricted analyses to the 500 bp regions surrounding each peak maxima. We computed the percentage of peaks that contained $\geq$1 motif site, and compared it to the percentage of negative control peaks that contained $\geq$1 motif site. The enrichment ratios were shown in Figure 3. The results show that TileProbe (TPV) again performed as well as or better than MAT. The greatest advantage was for conditions in which there was no control sample. Unlike the consistency test, TPV now outperformed MAT even in scenarios where control samples were available (e.g. 1IP 1CT and 3IP 3CT for Gli3, and 1IP 1CT for other TFs).

## 4.3 Comparison with other approaches

Next, we compared TPV with other variants of TileProbe and other data preprocessing approaches. In all comparisons, after data normalization and background correction, peaks were detected using the same peak detection procedure used for MAT and TileProbe.

Compared to TPM, TPV did not show noticeable advantage in the consistency test (Fig. 2). However, one cannot rule out the possibility that this is due to the gold standard bias, since the gold standard was constructed using MAT; and by dividing $\tau_i$, TPV has more procedural difference compared to MAT than TPM-MAT difference. In the motif enrichment test (Fig. 3), TPV performed better than TPM, suggesting that the variance correction in formula (7) can indeed help. In both tests, TPM performed comparable to or better than MAT.

Quantile normalization (QN) cannot be applied to peak detection without control samples, as illustrated in Figure 1a. However, it can

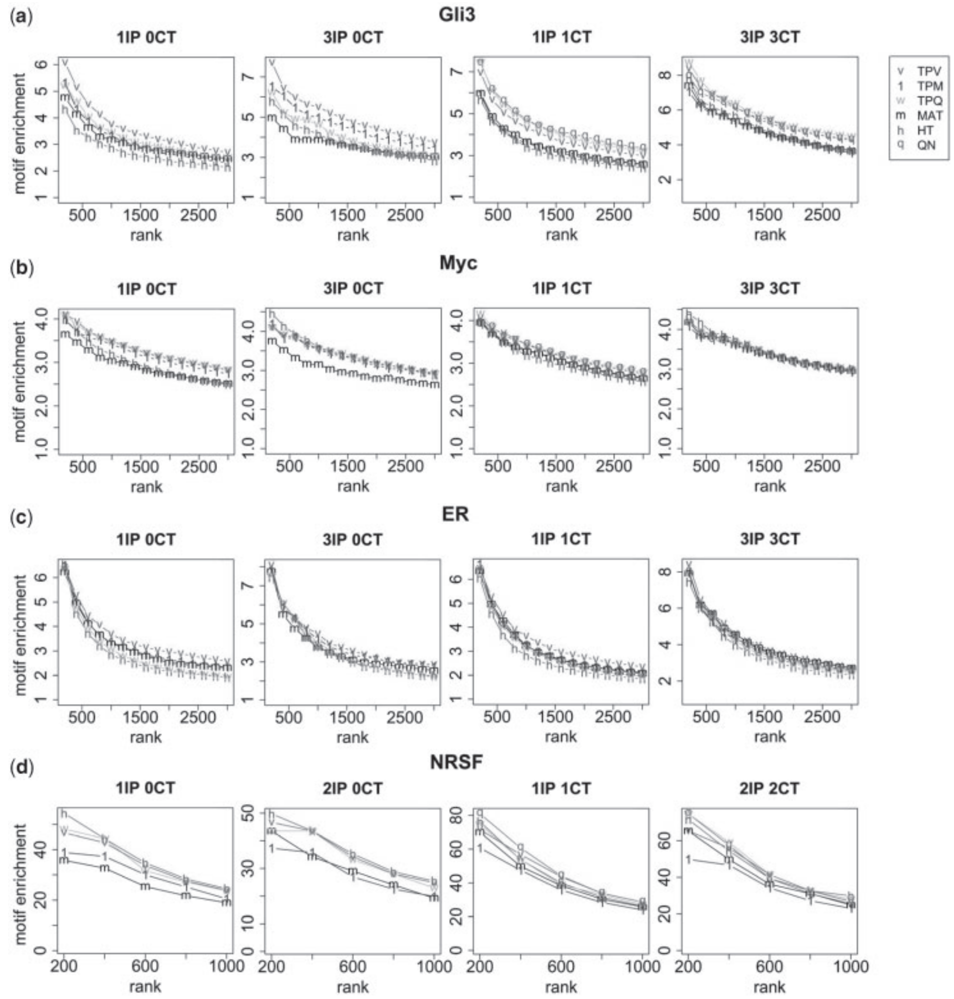

**Fig. 3.** Motif enrichment test. The enrichment ratios of the relevant transcription factor binding motif among the top 200, 400, 600, … , etc. peaks were shown for TileProbe-TPV (TPV), TileProbe-TPM (TPM), TileProbe-TPQ (TPQ), MAT, and HMMTiling (HT). For 1IP 1CT and 3IP 3CT (or 2IP 2CT for NRSF), the enrichment ratio was also shown for quantile normalization (QN). To avoid bias caused by peak length, all peaks were forced to be 500 bp long around the peak maxima.

be used in the 1IP 1CT and 3IP 3CT (or 2IP 2CT) analyses. The consistency test based on the MAT gold standard (Fig. 2) shows that, compared to TPV and MAT, QN performed slightly better in the Gli3 1IP 1CT analysis, slightly worse in NRSF 1IP 1CT, and comparable in Myc and ER 1IP 1CT. When we replaced the MAT gold standard by a gold standard constructed from the fullest QN analysis (i.e. 3IP 3CT or 2IP 2CT), QN performed better than or as well as TPV and MAT in all 1IP 1CT analyses (Fig. S3). In the motif enrichment test (Fig. 3), QN slightly outperformed TPV in the Gli3 1IP 1CT analysis, but TPV slightly outperformed QN in ER 1IP 1CT. Both performed better than or comparable to MAT in all four data sets. To summarize, in all three tests, QN slightly outperformed TPV in the Gli3 1IP 1CT analysis. In other scenarios, these two methods performed similarly, and no one consistently won the other. While TPV consistently performed better than or as well as MAT, QN did not: it performed worse in NRSF 1IP 1CT in the consistency test based on the MAT gold standard, even though this is likely due to gold standard bias. Among MAT, QN and TPV, only TPV is applicable to and performed best or among the best in all analytical conditions, including those without control samples.

Although QN cannot be applied without control samples, a probe effect model based on QN (i.e. TileProbe-TPQ) can be built using publicly available data, which can then be used to analyze new data even without controls. Both the consistency tests (Figs 2 and S3) and motif enrichment test (Fig. 3) show that TPQ did not perform as well as TPV (except for Gli3 1IP 1CT where TPQ performed slightly better). Although TPQ performed better than MAT in many cases (e.g. Gli3 and Myc 1IP 0CT, 3IP 0CT), it did not consistently do so. For example, it performed worse with respect to motif enrichment in the ER 1IP 0CT analysis (Fig. 3). In contrast, TPV, which was built upon MAT, consistently performed better than or equal to MAT in all analyses. This indicates that building TileProbe on MAT (rather than QN) is necessary to consistently gain over MAT.

In the HMMTiling (HT) approach proposed by Li *et al*. (2005), a background probe model was built empirically using multiple data sets generated by a single lab. The model was used as the emission probability of the null state of a hidden Markov model for peak detection. This method assumes that background behavior of probe $i$ follows normal distribution $N(\theta_i, \tau_i^2)$. To estimate $\theta_i$ and $\tau_i^2$, one first excludes the top 0.5% probes with the highest fluorescence intensities from each ChIP sample. Using the remaining 99.5% of the data points from the ChIP samples and all data points from the control samples, $\theta_i$ and $\tau_i$ are estimated using the sample mean and sample standard deviation of the quantile normalized log probe intensities at each probe. In principle, this method could be generalized to correct for probe effects after incorporating data from multiple labs and multiple studies. We tested this idea via replacing the $\theta_i$ and $\tau_i$ in formula (7) used by TPQ by the $\theta_i$ and $\tau_i$ computed using HT. The consistency tests (Figs 2 and S3) and motif enrichment test (Fig. 3) show that TPQ performed as well as or better than HT in most data sets expect for NRSF; neither MAT nor HT consistently outperformed the other; and TPV consistently performed better than or comparable to HT. There are three differences between TPQ and HT: (i) TPQ uses median instead of mean to estimate $\theta_i$; (ii) TPQ uses a shrinkage estimator in the variance estimates whereas HT uses sample variance directly; (iii) TPQ first computes sample variance within each experimental condition and then pools them together, whereas HT uses sample variance across all samples and ignores which dataset and experimental condition each sample comes from. Further tests showed that (iii) is the major factor that caused the difference in performance between TPQ and HT (see Supplementary Material S1).

## 4.4 How many samples are needed for building a robust probe effect model?

By randomly excluding an increasing amount of GEO data sets and array samples from the training data, we investigated the minimum number of samples necessary to robustly build the probe effect model in TileProbe. Using the probe models obtained from the reduced training data, Gli3 and ER were analyzed by TileProbe (TPV) again. The motif enrichment of peaks detected using different amounts of training data were compared (Figs 4 and S4). The results suggest that TileProbe performed consistently better than MAT, as long as the training data contained three or more independent studies and about 20 or more samples. When the model was built using only one study and six samples, the performance of TileProbe decreased significantly. In the most extreme case, when the probe model is built using a single study containing 4–6 samples and the training data involves the same transcription factor as the new data to be analyzed, signals in the new data may be subtracted away by the TileProbe background correction. This may result in bad performance of peak detection. Figure S5 provides such an example. Therefore, we recommend always using as many samples as possible and using samples from diverse studies and experimental conditions to build the TileProbe probe effect model. When there are <3 independent studies and <20 samples for building the model, TileProbe results should be interpreted with extra caution.

## 5 DISCUSSION

To summarize, we have proposed a new approach for removing probe effects from Affymetrix tiling array data. This approach takes advantage of hundreds of array samples already stored in public databases. Although the model we used is simple, our tests on different ChIP-chip data sets show that it is robust and performed surprisingly well compared to MAT. In certain analytical conditions, 30% gain was observed.

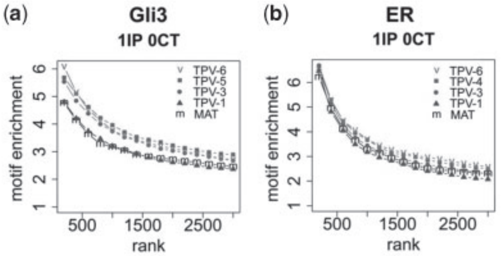

**Fig. 4.** Motif enrichment after reducing the number of samples used for building probe model. The enrichment ratios of the relevant transcription factor binding motif among the top 200, 400, … , etc. peaks were shown for TileProbe-TPV and MAT. (**a**) Gli3, Affymetrix Mouse Promoter 1.0R Array; TPV-6: TileProbe probe model trained using six independent studies (75 samples); TPV-5: five independent studies (38 samples); TPV-3: three studies (19 samples); TPV-1: one study (six samples). (**b**) Estrogen receptor, Affymetrix Human Tiling 2.0R Array 6; TPV-6: model trained using six independent studies (126 samples); TPV-4: four studies (48 samples); TPV-3: three studies (19 samples); TPV-1: one study (six samples). Only 1IP 0CT analyses are shown. Results for 1IP 1CT, 3IP 0CT and 3IP 3CT can be found in Figure S4.

In principle, the idea behind TileProbe is similar to the one used by Zilliox and Irizarry (2007) where thousands of gene expression microarray samples were used to construct a bar code for predicting the tissue origin of a new microarray sample. In a previous study, Huber *et al*. (2006) developed a data normalization procedure for transcriptome tiling array analysis. This represents another method that tries to remove probe effects empirically using the control samples. This method was not compared here as it was developed for yeast transcriptome analysis and is not directly applicable to our ChIP-chip data. It would be interesting to investigate in the future whether this approach can be extended to ChIP-chip, and if so, whether it can be tailored to use data from multiple studies (similar to tailoring QN to TPQ), and how it performs.

Our comparison between TPV and TPM indicates that probe-specific variability $\tau_i$ contributes to the improved performance of TileProbe. Although we did not encounter any data sets where TPM convincingly outperforms TPV, we cannot rule out the possibility that TPM may outperform TPV in future analyses. We speculate that if these cases exist, it will most likely occur when there are only limited amount of training data for building the probe model such that the variance estimates $\tau_i$ are heavily influenced by characteristics in the training data that do not generalize well to new data. In such cases, performance of TPV may be affected, and one may want to use TPM instead as it still outperforms MAT. To judge whether $\tau_i$ estimates are problematic, one may develop a sampling based method (see Supplementary Material S2). However, a more direct approach to decide whether one should use TPV or TPM is to apply both to the data of primary interest, and compare them using independent sources of information such as motif enrichment, correlation with gene expression changes, or qPCR validation rates. Such information is routinely generated in conjunction with ChIP-chip and should be available in most studies.

The current study represents the first one that systematically compares multiple data normalization procedures in the context of tiling array analysis. Our results show the advantage of the multi-sample driven strategy TileProbe over the single-sample driven method MAT. MAT and quantile normalization are two of the most popular methods currently used in the ChIP-chip community. Previous comparisons of ChIP-chip analysis algorithms based on these different preprocessing techniques only compared the end results of different tools. These comparisons did not separate differences due to the use of different data preprocessing procedures from differences due to the use of different peak calling procedures, and both these differences were not separated from differences caused by using different probe mapping library files (i.e. BPMAP files). As a result, different data preprocessing techniques have never been directly compared to each other. In our current study, the same peak calling procedure and the same BPMAP files were used for all comparisons. Since these confounding factors were carefully controlled, the observed differences were attributed purely to the differences in background correction. To the best of our knowledge, this is the first comparison of this type documented in the tiling array literature.

In the current work, we only tested the TileProbe model using ChIP-chip data, mainly because it allows us to directly compare with MAT, which was designed for this analysis. ChIP-chip was chosen also because we can use the independent motif information to objectively evaluate performance. It is not always easy to obtain an unbiased gold standard list big enough to evaluate high throughput analysis tools. We note, however, that the MAT model can be applied to other types of data and other types of arrays. For example, Kapur *et al*. (2007) used MAT to remove probe effects from Affymetrix exon arrays. It is therefore possible that TileProbe could also be applied to other contexts, provided that most samples used for training the probe model do not contain relevant biological signals. The current implementation of TileProbe is applied to Affymetrix tiling arrays. We speculate that the same concept can potentially be generalized to other array platforms such as NimbleGen and Agilent, although the model may need to be tailored to accommodate longer probes and correlation between channels, which may not be trivial. NimbleGen and Agilent offer custom arrays. For these arrays, the TileProbe concept may not be suitable unless there are enough data accumulated in public database.

With the rapid development of high throughput sequencing technologies, many applications of tiling arrays can now find their counterparts based on the next generation sequencing (e.g. ChIP-chip versus ChIP-seq) (reviewed by Shendure and Ji, 2008). In the near future, however, tiling arrays will remain to be an important tool for various genome-wide studies due to its relatively low cost and relatively mature protocols. Indeed, new tiling array data sets are flowing into the GEO database every month. In this context, there is continuing need for gaining better understanding of the current data processing techniques and developing better methods for data analysis. TileProbe represents one such effort. More importantly, as the amount of data in the public databases increases, there is increasing need to integrate information across multiple data sets. Such integration often requires reexamination and reanalysis of the existing data. In addition to offering an analytical tool for this purpose, TileProbe provides an example illustrating that pooling the huge amount of information stored in public databases improves our understanding and interpretation of the data.

## ACKNOWLEDGEMENTS

The authors thank Rafael Irizarry for helpful discussions.

*Funding*: National Institute of Health [R01GM083084, T32MH14592], and the Johns Hopkins faculty professional development fund.

*Conflict of Interest*: none declared.

## REFERENCES

Barrett,T. *et al*. (2007) NCBI GEO: Mining tens of millions of ex-pression profiles—database and tools update. *Nucleic Acids Res.*, **35**, D760–D765.

Bernstein,B.E. *et al*. (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, **125**, 315–326.

Bertone,P. *et al*. (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306**, 2242–2246.

Bolstad,B.M. *et al*. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Carroll,J.S. *et al*. (2005) Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell*, **122**, 33–43.

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.

Huber,W. *et al*. (2006) Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics*, **22**, 1963–1970.

Irizarry,R.A. *et al*. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

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

Ji,H. *et al*. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.*, **26**, 1293–1300.

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

Kampa,D. *et al*. (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.*, **14**, 331–342.

Kapranov,P. *et al*. (2002) Large-scale transcriptional activity in chromosomes 21 and 22. *Science*, **296**, 916–919.

Kapur,K. *et al*. (2007) Exon array assessment of gene expression. *Genome Biol.*, **8**, R82.

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

Li,C. and Wong,W.H. (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.

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.

Liu,X.S. (2007) Getting started in tiling microarray analysis. *PLoS Comput. Biol.*, **3**, e183.

Ozsolak,F. *et al*. (2007) High-throughput mapping of the chromatin structure of human promoters. *Nat. Biotechnol.*, **25**, 244–248.

Shendure,J. and Ji,H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135–1145.

Urban,A.E. *et al*. (2006) High-resolution mapping of DNA copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **103**, 4534–4539.

Weber,M. *et al*. (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.*, **39**, 457–466.

Wu,Z. *et al*. (2004) A model based background adjustement for oligonucleotide expression arrays. *J. Am. Stat. Assoc.*, **99**, 909–917.

Yuan,G.C. *et al*. (2005) Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science*, **309**, 626–630.

Zhang,X. *et al*. (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. *Cell*, **126**, 1189–1201.

Zilliox,M.J. and Irizarry,R.A. (2007) A gene expression bar code for microarray data. *Nat. Methods*, **4**, 911–913.