BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 18 2005, pages 3629–3636
doi:10.1093/bioinformatics/bti593

Gene expression

TileMap: create chromosomal map of tiling array hybridizations

Hongkai Ji[1] and Wing Hung Wong[2,*]

[1]Department of Statistics, Harvard University, Cambridge, MA 02138, USA and [2]Department of Statistics, Stanford University, Stanford, CA 94305, USA

Received on May 2, 2005; revised on July 19, 2005; accepted on July 20, 2005
Advance Access publication July 26, 2005

ABSTRACT

Motivation: Tiling array is a new type of microarray that can be used to survey genomic transcriptional activities and transcription factor binding sites at high resolution. The goal of this paper is to develop effective statistical tools to identify genomic loci that show transcriptional or protein binding patterns of interest.

Results: A two-step approach is proposed and is implemented in TileMap. In the first step, a test-statistic is computed for each probe based on a hierarchical empirical Bayes model. In the second step, the test-statistics of probes within a genomic region are used to infer whether the region is of interest or not. Hierarchical empirical Bayes model shrinks variance estimates and increases sensitivity of the analysis. It allows complex multiple sample comparisons that are essential for the study of temporal and spatial patterns of hybridization across different experimental conditions. Neighboring probes are combined through a moving average method (MA) or a hidden Markov model (HMM). Unbalanced mixture subtraction is proposed to provide approximate estimates of false discovery rate for MA and model parameters for HMM.

Availability: TileMap is freely available at http://biogibbs.stanford.edu/∼jihk/TileMap/index.htm

Contact: whwong@stanford.edu

Supplementary information: http://biogibbs.stanford.edu/∼jihk/TileMap/index.htm (includes coloured versions of all figures)

1 INTRODUCTION

Tiling array is a new type of microarray that interrogates genomes with high-density probes. In a typical tiling array, probes are distributed along chromosomes approximately evenly at a density of one probe per 10–100 bp. When hybridized with RNA or chromatin immunoprecipitation (ChIP) samples, the array will detect genomic loci that show transcriptional or transcription factor binding patterns of interest (Kapranov *et al*., 2002, 2003; Cawley *et al*., 2004; Kampa *et al*., 2004). Owing to the high density of the probes, a whole genome can be surveyed in an unbiased manner at a high resolution.

Identifying genomic loci that show transcriptional or protein binding patterns of interest is a key step of digesting information from tiling array experiments. Currently, available tools to fulfill this task are few. Examples include G-TRANS (Kampa *et al*., 2004), a moving average (MA) method proposed by Keles *et al*. (2004) and a hidden Markov model (HMM) method proposed by Li *et al*. (2005). The available tools are not sufficient to meet the diversified needs of the biology community. For example, current tools mainly rely on one-sample and two-sample comparisons. However, in order to study a complex developmental process, one may need to do tiling array experiment under a number of different developmental stages and identify genomic loci with specific temporal or spatial transcriptional or transcription factor binding patterns. This will inevitably involve sophisticated multiple-sample comparisons that current tools cannot handle. Moreover, if one wishes to do experiment under multiple conditions, the number of replicates within each condition will be small owing to the cost constraint. How to make efficient use of the small number of replicates was not specifically considered in previous works.

The goal of this paper is to develop effective statistical models and algorithms to detect genomic loci that show hybridization patterns of interest. We will emphasize the tool's ability to do flexible multiple sample comparisons and to make efficient use of a small number of replicates. A two-step approach, TileMap, is proposed. In the first step, a test-statistic is computed for each probe based on a hierarchical empirical Bayes model. In the second step, test-statistics of probes within a genomic region are combined to infer whether the region has the hybridization pattern of interest or not. Hierarchical empirical Bayes model shrinks variance estimates and increases sensitivity of the analysis when the number of replicates is small. It also provides a flexible way to do complex multiple sample comparisons. Two different methods—an MA method and an HMM—are used to combine neighboring probes. Unbalanced mixture subtraction (UMS) is proposed to provide an approximate estimate of local false discovery rate for MA and model parameters for HMM. Cawley *et al*.'s (2004) chromosome 21 and 22 ChIP-chip experiment data are used to test and illustrate TileMap, where it shows improved performance over existing methods.

The moving average method was initially used by Keles *et al*. (2004) in analyzing tiling array data. In addition to the ability to do multiple sample comparisons, there are two main differences between TileMap and Keles's method. First, to compute a probe level test-statistic, Keles's method uses data only from the probe in question, whereas TileMap pools information from all the probes in the array via a closed-form empirical Bayes shrinkage estimator of variance. Recent studies showed that pooling information from all probes is an effective way to increase the sensitivity of gene selection

*To whom correspondence should be addressed.

© The Author 2005. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oupjournals.org

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use, please contact journals.permissions@oupjournals.org

from microarray experiment when the number of replicates in the experiment is small (Baldi and Long, 2001; Newton *et al*., 2004; Smyth, 2004), and variance is the main component through which information pooling takes effect (H. Ji and W. H. Wong, submitted for publication). TileMap applies this idea to tiling array analysis. Second, a different strategy is adopted by TileMap to determine the cutoff for making rejections. Keles's method uses bootstrap to estimate the null distribution of their 'scan-statistics' for choosing the cutoff. They made an implicit equal mean assumption, i.e. under the null hypothesis H_0, mean hybridization intensities are equal under different experimental conditions. Although this assumption may be reasonable for two-sample comparisons with H_0: $\mu_1 = \mu_2$, it is often inappropriate for false discovery rate (FDR) estimation when H_0 contains some random effects, e.g. H_0: $\mu_1 - \mu_2 \sim N(0,1)$, and for FDR control of multiple-sample comparisons, such as mutant1 (mt1) < wild type(wt) < mutant2 (mt2). For the latter case, the correct null hypothesis is H_0: NOT {mt1 < wt < mt2} instead of H_0': mt1 = wt = mt2. H_0 not only includes H_0', but also mt1 = wt < mt2, mt1 > wt < mt2, etc. FDR control for such a complicated composite null is difficult. To deal with this problem, TileMap adopts an empirical technique, i.e. UMS, to get an approximate estimate of the local FDR and to choose a cutoff. In contrast to TileMap and Keles's method, Affymetrix's G-TRANS uses a different strategy. In G-TRANS, probes are grouped into overlapping windows, and a Wilcoxon signed-rank or rank sum test is carried out for each window. It is difficult to generalize this method to complex multiple sample comparisons, and no FDR estimate is provided by G-TRANS. Recently, Li *et al*. (2005) also proposed an HMM method for tiling array analysis, but their method was again limited to two-sample comparisons and did not pool information across probes when estimating the variance of individual probes.

2 METHODS

2.1 Hierarchical empirical Bayes model for computing probe level test-statistics

After proper preprocessing of the data, the first step of TileMap is to compute a test-statistic for each probe. Assume that there are I probes in the array, hybridizations are done in J different conditions, and there are K_j replicates under condition j. Let X_{ijk} denote the normalized and log-transformed PM or PM-MM value of probe i under condition j and replicate k. The following model is used to describe X_{ijk}:

$$X_{ijk}|\mu_{ij},\sigma_i^2 \sim N(\mu_{ij},\sigma_i^2), \quad i = 1, 2, \ldots, I; \quad j = 1, 2, \ldots, J; \quad k = 1, 2, \ldots, K_j. \tag{1}$$

$$\mu_{ij}|\mu_0,\tau_0^2 \propto 1, \tag{2}$$

$$\sigma_i^2|v_0,\omega_0^2 \sim \text{Inv}-\chi^2(v_0,\omega_0^2). \tag{3}$$

Define $v = \Sigma_j(K_j - 1)$, $s_i^2 = \sum_j \sum_k (x_{ijk} - \bar{x}_{ij})^2/v$, $\overline{s^2} = \sum_i s_i^2/I$ and $S = \sum_i [s_i^2 - (\overline{s^2})]^2$. The basic idea to compute probe level statistics is to estimate σ_i^2s by pooling information from all s_i^2s, then treat $\sigma_i^2 = \hat{\sigma}_i^2$ as known and compare μ_{ij}s in terms of their posterior distribution.

To estimate σ_i^2, we use the following empirical Bayes shrinkage estimator proposed by H. Ji and W. H. Wong (submitted for publication), based on the theory of Natural Exponential Family with Quadratic Variance Function (Morris, 1983):

$$\hat{\sigma}_i^2 = (1 - \hat{B})s_i^2 + \hat{B}\overline{s^2}, \tag{4}$$

$$\hat{B} = \frac{2/v}{1 + 2/v}\frac{I-1}{I} + \frac{1}{1+2/v}\left(\frac{2}{v}\right)(\overline{s^2})^2\frac{I-1}{S}. \tag{5}$$

The derivation of the estimator is outlined in Section 1, Supplementary materials.

Once σ_i^2 is obtained, the posterior distribution of μ_{ij} will be approximated by $N(\overline{X}_{ij}, \hat{\sigma}_i^2/K_j)$, and probe level statistics will be constructed based on this approximation. For two-sample comparisons ($J = 2$), the probe level test-statistic is

$$t_i = \frac{\overline{X}_{i1} - \overline{X}_{i2}}{\hat{\sigma}_1\sqrt{(1/K_1) + (1/K_2)}}. \tag{6}$$

For multiple-sample comparisons ($J > 2$), e.g. {(m1 > wt) or (m2 > wt)}, the probe level test-statistics will be computed as follows: (1) draw μ_{ij}s from $N(\overline{X}_{ij}, \hat{\sigma}_i^2/K_j)$ C times; (2) for each probe i, count how many times the prespecified condition is satisfied, and denote this number by S_i; (3) use $t_i = 1 - (S_i/C)$ as probe level summary. The advantage of this simulation based method is its flexibility, making it especially useful to study, i.e. hybridizations at specific time and specific place in developmental processes.

Formula (6) above looks like a t-statistic, but, in fact, they are different in the way the denominator is derived. $\hat{\sigma}_i$ in formula (6) pools information from all probes, whereas s_i which is used in canonical t-statistics uses only information of probe i to estimate its own standard deviation. Pooling information to estimate variance significantly increases the sensitivity of the method, since it provides better estimate of variance in terms of mean square error and results in better separation of test-statistic distributions under the null and alternative hypothesis. The same principle applies to multiple-sample comparisons too. We also tried to shrink μ_{ij}s by setting a proper prior in formula (2). However, mean shrinking usually does not provide much additional gain in sensitivity whereas it may incur a significant amount of extra computation. This explains why a flat prior was used in formula (2). In formula (1), we assume common variance under different conditions, but this assumption is not crucial. One can assume unequal variance and apply the shrinkage estimator for each condition separately.

Without loss of generality, in what follows, we assume that small t_i corresponds to the hybridization pattern of interest. Depending on individual studies, this can be met by setting appropriate group labels (e.g. in a ChIP-chip experiment, define $\overline{X}_{i1}$ and $\overline{X}_{i2}$ in formula (6) to be the control and IP samples respectively) or by taking transformations, such as $-t_i$ if necessary.

2.2 Combining information from neighboring probes

TileMap provides two ways to combine information from neighboring probes. The first method uses an MA. In other words,

$$m_i = \frac{\sum_{k=i-w}^{i+w} t_k}{(2w+1)} \tag{7}$$

is computed as the final summary statistic for probe i. This is identical to Keles *et al*.'s (2004) scan-statistic, except for the way t_i is calculated. Keles's method uses canonical t-statistics, whereas here we use a modified version of it and pool information from all probes to estimate variance. Keles's method only considers two-sample comparisons, whereas TileMap can handle multiple-sample comparisons. Before taking average, t_i in multiple-sample case will be transformed by $\log[t_i/(1-t_i)]$. Notice that in multiple-sample case, t_i is a posterior probability and belongs to [0, 1]; if $t_i = 0$ or 1, it is replaced by ε or $1-\varepsilon$, where ε is a small number (e.g. 1×10^{-6}). For two-sample case, t_i is given by formula (6) and belongs to $(-\infty, +\infty)$, and it will be used directly in formula (7). The choice of w was discussed in Keles *et al*. (2004) and will not be the focus here.

The second method uses HMM. The advantage of using HMM is that there is no need to preselect a w before analyzing new data. The HMM structure is shown in Figure 1b. More precisely, let H_i denote the hybridization state of probe i. $H_i = 1$ if probe i shows the pattern of interest; otherwise $H_i = 0$. Hereafter, we assume that $i = 1, \ldots, I$ correspond to the probe's physical order on the chromosome. Define $d_{i,j}$ to be the physical distance between the centers of probes i and j. Assume that (1) $P(H_i = 0) = \pi_0, P(H_i = 1) = \pi_1 = 1 - \pi_0$; (2) if $d_{i,i+1} \le d_0$, the transition probabilities are $P(H_{i+1} = 1|H_i = 0, d_{i,i+1} \le d_0) = a_0$,

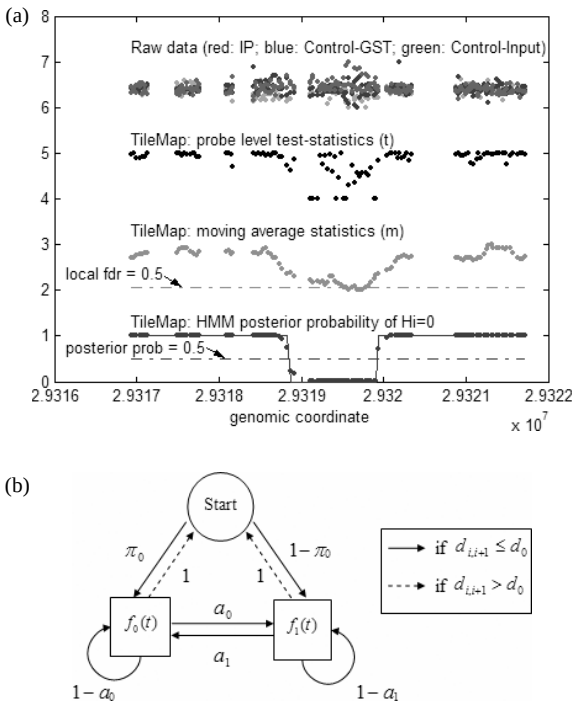

Fig. 1. TileMap overview. (**a**) Illustration of TileMap procedures. Raw data, TileMap probe level test-statistics, MA summaries and HMM posterior probabilities are shown from top to bottom. In TileMap, small test-statistics correspond to the hybridization pattern of interest. The posterior probability shown is the posterior probability of not being a target probe. (**b**) The HMM structure in TileMap.

$P(H_{i+1} = 0|H_i = 1, d_{i,i+1} \le d_0) = a_1$; if $d_{i,i+1} > d_0$, $P(H_{i+1} = 0|H_i, d_{i,i+1} > d_0) = \pi_0$, $P(H_{i+1} = 1|H_i, d_{i,i+1} > d_0) = \pi_1$; (3) the conditional distribution of probe level test-statistics is $f(t_i = t|H_i = 0) = f_0(t)$, $f(t_i = t|H_i = 1) = f_1(t)$. Under these assumptions, once d_0, π_0, a_0, a_1, $f_0(t)$ and $f_1(t)$ are known, the standard forward–backward algorithm can be applied to infer the hidden state H_i through probe level test-statistics t_i.

In MA, m_is are used to rank and select probes to form target regions. The entire set of m_i can be viewed as a sample from a mixture distribution $\pi_0 f_0(m) + \pi_1 f_1(m)$, where $f_0(m)$ and $f_1(m)$ are distributions of m_i under $H_i = 0$ and $H_i = 1$, respectively. We need to estimate π_0, $f_0(m)$ and $f_1(m)$ in order to control FDR. In HMM, t_is are used to infer the hidden states, and target regions are selected based on the posterior probability of $H_i = 1$. The t_is can be treated as a sample from another mixture $\pi_0 f_0(t) + \pi_1 f_1(t)$, and one needs to know π_0, $f_0(t)$ and $f_1(t)$ before decoding the HMM. TileMap adopts unbalanced mixture subtraction to deal with these two similar issues.

2.3 UMS

The goal of UMS is to recover different components of a mixture distribution $h(t) \equiv \pi_0 f_0(t) + (1 - \pi_0) f_1(t)$, where t represents a generic statistic. In reality, $h(t)$ is observed, but π_0 and $f_1(t)$ are unknown. Canonical FDR procedures assume $f_0(t)$ to be known and try to restore $f_1(t)$ by subtracting $f_0(t)$ from $h(t)$. These procedures usually work well in two-sample comparisons where $f_0(t)$ can be obtained, either by theory or by simulation techniques, such as permutations. However, they cannot be applied directly to cases where $f_0(t)$ is hard to obtain, e.g. multiple-sample comparisons or comparisons involving complex composite null. To circumvent this difficulty, UMS makes use of additional information (see below) to construct two 'unbalanced' distributions. Both are mixtures of $f_0(t)$ and $f_1(t)$, but they differ in

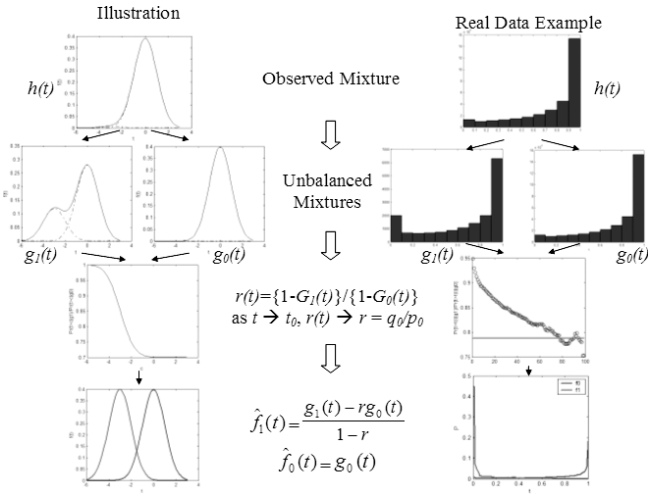

Fig. 2. Unbalanced Mixture Subtraction. Left panel is a conceptual example to illustrate UMS. See Section 2.3 for a detailed description. Right panel is a real example where UMS was applied to analyze 18 arrays of a cMyc ChIP-chip experiment to estimate $f_0(t)$ and $f_1(t)$ (Section 4).

the abundance of $f_1(t)$ component. The two 'unbalanced' mixtures can then be used to reconstruct π_0, $f_0(t)$ and $f_1(t)$ and estimate FDR.

UMS is illustrated in Figure 2. We first construct two mixtures $g_0(t) = p_0 f_0(t) + (1 - p_0) f_1(t)$ and $g_1(t) = q_0 f_0(t) + (1 - q_0) f_1(t)$, where $p_0 > \pi_0 \ge q_0$. If two such mixtures can be constructed and if $\exists t_0$ such that $t \to t_0$, $f_0(t)/f_1(t) \to \infty$, then $\lim_{t \to t_0} g_1(t)/g_0(t) = q_0/p_0 \equiv r$. Once r is known, $f_1(t)$ can be obtained by formula (8).

$$f_1(t) = \frac{g_1(t) - r g_0(t)}{1 - r}. \tag{8}$$

To estimate $f_0(t)$, notice that $\pi_1 = 1 - \pi_0$ is usually small; therefore, $g_0(t)$ can provide an approximation of $f_0(t)$. Given $f_1(t)$ and $g_0(t)$, π_0 can be estimated by fitting $h(t)$ using $\theta_0 g_0(t) + (1 - \theta_0) f_1(t)$ such that $\int \{h(t) - [\theta_0 g_0(t) + (1 - \theta_0) f_1(t)]\}^2 \, dt$ is minimized. The resulting estimate $\hat{\theta}_0 = \{\int [h(t) - f_1(t)][g_0(t) - f_1(t)] \, dt\} / \{\int [g_0(t) - f_1(t)]^2 \, dt\} = \pi_0/p_0 \ge \pi_0$; therefore, $\hat{\pi}_1 = 1 - \hat{\theta}_0$ provides a conservative estimate of π_1, which is desirable if we want to keep a relatively stringent criteria in detecting regions of interest. Once π_1, $f_0(t)$ and $f_1(t)$ are estimated, local false discovery rate at a point t can be estimated by $\text{lfdr}(t) = \hat{\pi}_0 \hat{f}_0(t)/[\hat{\pi}_0 \hat{f}_0(t) + \hat{\pi}_1 \hat{f}_1(t)]$, and FDR for a rejection region Z, e.g. $\{t \le t_{\text{cut}}\}$, can be estimated using the relationship $\text{FDR}(Z) = \text{E}[\text{lfdr}(t)|Z]$. In the special case where the null distribution $f_0(t)$ is known, we can set $p_0 = 1$, $g_0(t) = f_0(t)$, and $g_1(t) = h(t)$; then UMS reduces to the q-value method discussed in Storey and Tibshirani (2003).

To construct the two unbalanced mixtures $g_0(t)$ and$g_1(t)$, we need additional information. If biological knowledge tells that certain regions are more likely to be transcribed or bound by the transcription factors under study, this piece of information can be used, e.g. one can map known transcription factor binding motifs to the genome to collect regions of potential interest. If such biological information is not available, the correlation structure provided by tiling array itself can be used to get an approximate estimate of $g_0(t)$ and $g_1(t)$ as discussed in the following paragraphs.

For HMM, we pick up probes with $t_i > t_{(p)}$, where $t_{(p)}$ is the p-th percentile of all t_is. Then, t_{i+1}, the immediate downstream test-statistics of the selected probes are used to form $\tilde{g}_0(t)$. We then pick up probes with $t_i \le t_{(q)}$, and use their downstream t_{i+1} to form $\tilde{g}_1(t)$. For MA, t_{i+1} is replaced by m_{i+w+1}, and a similar procedure is used to construct $\tilde{g}_0(m)$ and $\tilde{g}_1(m)$. We then use $\tilde{g}_0(.)$ and $\tilde{g}_1(.)$ to surrogate $g_0(.)$ and $g_1(.)$. The intuition behind these procedures is that if a DNA/cDNA fragment hybridizes to a probe, it also tends to hybridize to its neighboring probes. Thus, if a probe has very

small t_i, its neighboring probes are more likely to have the pattern of interest than random probes do and vice versa.

To generalize the above procedures, we define a selection statistic u_i. We use $\tilde{g}_0(t) = f(T_i = t|\, I_{\{u_i \in A\}} = 1)$ to approximate $g_0(t)$, and $\tilde{g}_1(t) = f(T_i = t|I_{\{u_i \in R\}} = 1)$ to approximate $g_1(t)$. For MA, $u_i = t_{i-w-1}$, $T_i = m_i$. For HMM, $u_i = t_{i-1}$, $T_i = t_i$. Both MA and HMM use $A = \{u_i > t_{(p)}\}$ and $R = \{u_i \leq t_{(q)}\}$. By default, $t_{(p)} = t_{(1)}$ and $t_{(q)} = t_{(5)}$ (see Section 5.3 Supplementary material for discussions about the choice of $t_{(p)}$ and $t_{(q)}$). As an illustration, the right panel of Figure 2 provides a real example for estimating HMM parameters in this way.

It can be shown that if (1) $P(H_i = 0|I_{\{u_i \in A\}} = 1) > \pi_0$, (2) $f(T_i = t|H_i,\ I_{\{u_i \in A\}} = 1) = f(T_i = t|\, H_i)$, then $\tilde{g}_0(t)$ is a valid surrogate for $g_0(t)$. Similarly, if (1) $P(H_i = 0|I_{\{u_i \in R\}} = 1) \leq \pi_0$, (2) $f(T_i = t|H_i,\ I_{\{u_i \in R\}} = 1) = f(T_i = t|\, H_i)$, $\tilde{g}_1(t)$ is a valid surrogate for $g_1(t)$. Usually, condition (1) is not hard to meet. Condition (2) is implied in HMM case, but in general, it only holds approximately or may not even hold owing to the possible selection bias or the residual correlations between u_i and T_i after accounting for H_i. We, therefore, label the application of UMS here as an 'approximate' procedure, meaning that it only provides a rough and possibly imprecise or optimistic estimate of FDR under the null model, unless the previous assumptions are completely satisfied. The advantage of UMS over the permutation-based FDR estimation, such as SAM (Tusher *et al*., 2001), is that, if conditions (1) and (2) are indeed satisfied, UMS can provide FDR estimate for complex composite null hypothesis, such as 'not $\mu_1 < \mu_2 < \mu_3$ or $\mu_4 < \mu_5$', whereas the latter cannot. Also, UMS provides an interface to incorporate other sources of information (e.g. empirical biological knowledge about which genes/regions are more likely to show desired pattern) to evaluate false positive rates. When applying UMS, however, it is important to understand that there is always a possibility that bias may be introduced by the new sources of information.

For HMM, one also needs to determine a_0, a_1 and d_0. One can choose a_1 and d_0 according to the typical length of hybridizations. For example, in ChIP-chip experiments, IP fragments are usually ~1 kb. If the probe density in the chip is 1 probe/35 bp, a typical hybridization would contain ~28 probes; correspondingly, a_1 can be set to 1/28 to match the mean length of continuous $H_i = 1$ segments in HMM, and d_0 can be set to 1000. To estimate a_0, assume that (π_0, π_1) is the stationary distribution for the Markov chain without gaps (i.e. without $d_{i,i+1} > d_0$), then $\pi_1 = a_0/(a_0 + a_1)$, and a_0 can be estimated by $\hat{a}_1\hat{\pi}_1/(1 - \hat{\pi}_1)$ where $\hat{\pi}_1 = 1 - \hat{\theta}_0$.

3 IMPLEMENTATION

TileMap is implemented in ANSI C. In terms of computation time, it is usually >10 times faster than G-TRANS (refer to Section 2, Supplementary material). TileMap includes functions to do raw data normalization, local repeat filtering, probe level summary, UMS, MA and HMM. Local repeat filtering masks any probe that occurs more than once in a 2 kb local window. In UMS, users may choose to use their own selection statistics. For MA, permutation-based FDR estimation routine is also provided. The output of TileMap includes final summaries for each probe and a *.bed file containing selected genomic regions. The latter is defined by lfdr in MA or posterior probability of $H_i = 0$ in HMM being smaller than a user specified cutoff.

In UMS, all statistics are transformed to [0, 1], e.g. t-statistic is transformed using $\exp(t_i)/[1+\exp(t_i)]$. [0, 1] is then equally divided into n (default = 1000) intervals. $g_0(.)$ and $g_1(.)$ were estimated using empirical distributions of test-statistics in these intervals. To estimate r, we compute $r_t = [1 - G_1(t)]/[1 - G_0(t)]$ for $t = t_{(50)}, t_{(51)}, \ldots, t_{(99)}$. r is then set to be the median of these 50 r_ts. To get stable estimates of $f_1(.)$, we also assumed monotone likelihood ratio in implementing UMS, i.e. as $t \to t_0$, $f_0(t)/f_1(t)$ is increasing.

4 RESULTS

Tilemap was tested using a ChIP-chip experiment performed by Cawley *et al*. (2004) as well as simulations. In this section, we will present the global design and the main results of the tests. Details of how tests and simulations were done are provided in Supplementary material (Sections 3–6). Cawley's experiment tried to identify binding regions for three transcription factors using Affymetrix chromosomes 21 and 22 tiling arrays. Their cMyc data on Chip A and p53-FL (full length antibody) data on Chips A, B, C were used for testing. For discussion's convenience, Chips A, B and C in p53 experiment are treated here as a combined single chip. For each transcription factor, hybridizations were done for two biological replicates under three different conditions: IP, control GST (C1) and control input (C2). For each biological replicate and experimental condition, three technical replicates were obtained. In total, there were 18 arrays for each transcription factor. Before analysis, raw data were quantile normalized (Bolstad *et al*., 2003), PM-only intensities were log transformed and adjusted for batch effect (Section 3.1, Supplementary material). Local repeats were filtered out. The 18 arrays were then randomly divided into three groups G1, G2 and G3 for later use. Each group contained six arrays: two for IP, two for C1 and two for C2. Within each condition, the two arrays were from different biological replicates.

4.1 Sensitivity test based on cMyc data

In order to see how variance shrinking helps increase sensitivity in small replicate case, MA with variance shrinking (MA-S) was first compared with MA without variance shrinking (MA-N). Notice that in two sample comparisons, MA-N is equivalent to Keles's scan statistic. Before the comparison, a gold standard was constructed by applying MA-N to all 18 arrays to select probes showing IP > C1 and IP > C2 (Section 3.2, Supplementary material). The gold standard contained 1654 probes (0.5% of all probes) and was grouped into 180 binding regions. In order to compare, one or two of the G1–G3 groups were excluded. MA-S($w = 5$) and MA-N($w = 5$) were applied to the remaining arrays to rank probes according to (1) IP > C1 using one of the G1–G3 groups only; (2) IP > C1 and IP > C2 using one of the G1–G3 groups only; and (3) IP > C1 and IP > C2 using two of the G1–G3 groups. For simplicity, we use s2r2, s3r2 and s3r4 to denote the three settings above. s2r2 stands for two-sample comparison where each sample has two replicates, s3r2 stands for three-sample two replicates, etc. In each of the three settings above, 0.5% of top ranking probes were selected to form binding regions. This guaranteed that both methods had the same coverage of the genome. Two probes, if separated by <500 bp, were treated as in a single region. Regions were ranked according to minimum of m_is of each region. MA-S and MA-N were then compared in terms of what fraction of their top ranking probes were gold standard probes (Fig. 3a) and how many of their top ranking regions overlap gold standard regions (Fig. 3b). There were three possibilities to choose one group or two groups to exclude, and the results shown here were averages over the three possibilities. According to Figure 3, MA-S was indeed more powerful than MA-N, even if the way we define gold standard was biased toward MA-N. In s2r2 case, the effect was most striking. At probe level, the rate of correct rejections increased from ~0.2 to ~0.85 when 500 rejections were made; at binding region level, MA-S identified >70 more gold standard regions among the top 160 regions. The gain from shrinking decreases as the number of arrays

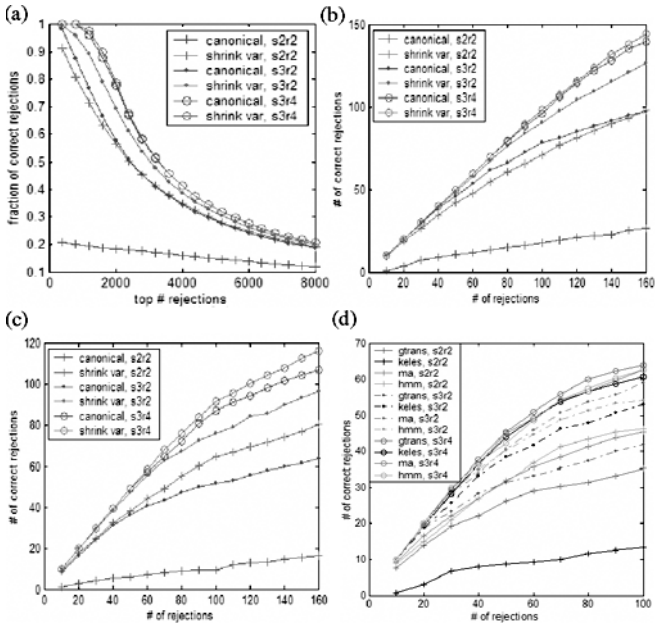

Fig. 3. Comparisons of MA, HMM, Keles's method and G-TRANS in cMyc data analysis. Fraction/number of correct rejections among certain number of total rejections was shown. (**a**) MA-S and MA-N were compared at probe level, probe density $= 1/35$ bp; (**b**) MA-S and MA-N were compared at binding region level, probe density $= 1/35$ bp; (**c**) MA-S and MA-N were compared at binding region level, probe density $= 1/70$ bp; (**d**) G-TRANS, Keles's method (MA-N), MA-S and HMM were compared at binding region level, probe density $= 1/35$ bp, G–M was used as gold standard.

increases, but given that 2–3 replicates are most often encountered, we would expect to gain from variance shrinking in a significant number of real studies.

We also reduced the probe density from 1 probe/35 bp to 1 probe/70 bp by discarding half of the probes. MA-S($w = 2$) and MA-N($w = 2$) were compared again using the data with the reduced probe density (Fig. 3c). The gold standard used in Figure 3c was the same as that in Figure 3b which were constructed using all probes including the probes discarded. The gain from variance shrinking became more evident. Interestingly, in s3r2 case, MA-S found ∼100 'true' binding regions among its top 160 rejections (Fig. 3c). The same sensitivity was achieved by MA-N but with doubled probe density (Fig. 3b). This means that we only need half as many probes for MA-S as for MA-N to achieve the same sensitivity in this case. If MA-N were used to survey 100 genes, using the same number of probes, MA-S allows us to survey 200 genes without losing the ability to detect the true targets. Surveying more genes, however, can increase the chance for finding the unknown players in a biological process.

In order to compare G-TRANS, Keles's method, MA-S and HMM, they were applied to analyze cMyc data as we did in Figure 3b. Two gold standards, 'G–M' and 'G–H', were constructed using all 18 arrays (Section 3.3, Supplementary material). G–M standard contained 78 regions, which is the intersection of the regions identified by G-TRANS and MA-N. G–H standard contained 73 regions, which is the intersection of G-TRANS and HMM regions. Different methods were compared at binding region level using both G–M standard (Fig. 3d) and G–H standard (Supplementary Figure S2). The two results were similar. When replicates were few (s2r2, s3r2), MA-S and HMM showed clear improvement in sensitivity compared with G-TRANS and Keles's method. As the number of arrays became large (s3r4), all methods began to show similar performance. Keles's method and G-TRANS cannot be used to do multiple sample comparisons. In order to get summary statistics for IP $>$ C1 and IP $>$ C2, Keles's method was replaced by MA-N; G-TRANS was applied twice to do two-sample comparisons IP $>$ C1 and IP $>$ C2 separately, and the maximum of the two P-values for each probe was taken as its final summary statistics to derive binding regions. We did not compare TileMap with the HMM method proposed by Li *et al*. (2005), since the latter was not available at the time this work was done.

Next, we compared the enrichment of cMyc binding sites in regions identified by different methods. cMyc consensus binding pattern CA[C/T]G[T/C]G was mapped to chromosome 21, yielding 17 563 potential binding sites (TFBS) in a total of 18.3 Mb non-repeat genomic sequences. Among these TFBS, 4496 were located in regions whose human–mouse–rat cross-species conservation score was among the top 30% of the whole chromosome (Section 3.4, Supplementary material). G-TRANS, MA-N (Keles), MA-S and HMM were all applied to select the top 0.5% probes and group them into binding regions using 18 cMyc arrays (s3r6) as well as reduced data (s2r2, s3r2, s3r4). The number of TFBS and conserved TFBS (cTFBS) in the identified regions were counted. Binding site enrichment was computed as the ratio of TFBS and cTFBS densities in selected regions to their chromosome-wide counterparts. Site enrichment for different methods is listed in Table 1. Based on the results, MA-S and HMM consistently showed higher or nearly equal TFBS and cTFBS enrichment than G-TRANS. They also showed higher TFBS enrichment than MA-N (Keles) in s2r2 case. The differences, however, diminished as more arrays were included.

4.2 Sensitivity test based on p53 data

Different methods were further compared through the analysis of 18 p53-FL arrays. Cawley *et al*. (2004) verified 14 p53 binding regions by qPCR, using either p53_FL or p53_DO1 antibody. These regions were used here as gold standard. Each method was applied under different settings (s2r2–s3r6) to select the top 0.5% probes and group them into binding regions. Methods were then compared in terms of how many experimentally verified regions were identified in their top 10, top 20 and all selected regions. The results were listed in the top panel of Table 2. HMM and MA-S again detected more experimentally verified regions than GTRANS and MA-N (Keles) when replicates were few (e.g. s2r2, s3r2). We also reduced the probe density from 1/35 to 1/100 bp by discarding two-thirds of all the probes. MA-N, MA-S and HMM were compared again (the bottom panel of Table 2). The better performance of MA-S and HMM over MA-N in small replicate case became more evident (e.g. s3r2). G-TRANS was not compared here since we were unable to use it to analyze a set of specified probes.

4.3 Performance of UMS

To see how UMS works, a series of simulations were done. In all simulations, six arrays were generated and equally divided into three groups D1, D2 and D3, each of size two. Each array contained 50 000 probes. Probe intensities were generated according to formulae (1)

Table 1. cMyc binding site enrichment in predicted binding regions

Method	s2r2	s3r2	s3r4	s3r6
GTRANS	1.2 (0.1)/1.1 (0.1)	1.4 (0.1)/1.1 (0.1)	1.8 (0.1)/1.4 (0.1)	1.9/1.5/96k
MA-N/Keles	1.6 (0.2)/1.1 (0.1)	1.9 (0.3)/1.4 (0.2)	2.0 (0.1)/1.5 (0.1)	2.0/1.5/150k
MA-S	1.7 (0.2)/1.3 (0.1)	1.9 (0.2)/1.4 (0.1)	2.0 (0.1)/1.5 (0.1)	2.0/1.5/149k
HMM	1.9 (0.2)/1.4 (0.1)	2.0 (0.2)/1.4 (0.1)	2.0 (0.1)/1.4 (0.1)	2.0/1.4/134k

For reduced data (s2r2, s3r2, s3r4), conserved TFBS (rc) and TFBS enrichment (rt) were shown as rc(se)/rt(se). rc and rt were averages over three independent analyses, se was standard error of the average. When all 18 arrays were analyzed (s3r6), se cannot be computed, the number of non-repeat bases nb in the predicted regions was shown instead. The results were in the format rc/rt/nb.

Table 2. Sensitivity of GTRANS, MA-N, MA-S and HMM on p53 data

Methods	s2r2	s3r2	s3r4	s3r6
GTRANS	3.3/4.7/8.7	4.3/8.0/11.7	5.0/9.3/12.3	6.0/10.0/12.0
MA-N/Keles	0.7/1.0/4.0	6.3/9.0/12.7	6.0/10.0/13.0	6.0/10.0/13.0
MA-S	6.0/10.0/13.0	6.7/10.0/12.7	6.0/10.0/13.0	6.0/10.0/13.0
HMM	7.0/9.7/11.3	6.7/9.0/12.3	6.3/9.7/13.0	7.0/10.0/13.0
MA-N/Keles/3	0.0/0.0/1.3	3.0/4.3/9.7	4.3/6.0/12.0	4.0/5.0/12.0
MA-S/3	3.3/6.0/11.0	3.7/6.0/11.0	4.7/5.3/12.0	4.0/5.0/12.0
HMM/3	4.0/5.0/9.0	5.0/6.3/10.7	4.3/6.7/10.7	5.0/7.0/11.0

Number of experimentally verified p53 regions among top 10 ($n1$), top 20 ($n2$) and all regions ($n3$) identified by different methods were shown as $n1/n2/n3$. For reduced data (s2r2, s3r2, s3r4), the numbers shown were averages over three analyses.

Table 3. Simulation design for testing UMS

Number	Within target region (H_i=1)	Outside target region (H_i=0)
I	Common for I–III:	$\Delta_{i1} = \Delta_{i2} = 0$
II	$\Delta_{i1} = \lvert r_{i1}\rvert, \Delta_{i2} = \lvert r_{i2}\rvert$ $r_{i1}, r_{i2} \sim N(1, 0.25)$	$\Delta_{i1}, \Delta_{i2} \sim N(0, 0.25)$ Δ_{ij}s are all independent
III	r_{ij}s are all independent	$\Delta_{i1} = \Delta_{i2} = 0$ plus two types of binding regions with other patterns: (a) $\Delta_{i1} = 0, \Delta_{i2} = \lvert r_{i2}\rvert$ (b) $\Delta_{i1} = \lvert r_{i1}\rvert, \Delta_{i2} = 0$ $r_{i1}, r_{i2} \sim N(1, 0.25)$ i.i.d

and (3). $v_0 = 4.64$, $\omega_0^2 = 0.021$ were chosen to match the typical values observed in real data. A number of binding regions with pattern D1 < D2 < D3 were generated to serve as targets, we wish to identify (Section 4.1, Supplementary material). In total, these regions covered 50 000 π_1 probes. Simulations differ in the way $\Delta_{i1} = \mu_{i2} - \mu_{i1}$ and $\Delta_{i2} = \mu_{i3} - \mu_{i2}$ were generated, which was designed to test UMS from different perspective. Table 3 listed the designs for simulations I–III. In each simulation, 10 different datasets were generated, and the results below were averages over the 10 datasets. Here, we use $\pi_1 = 0.05$ to illustrate the results, although $\pi_1 = 0.01$, 0.02 and 0.10 were also tried and similar results were obtained.

Simulations I and II tested UMS when its assumptions were true. In simulation I, we tested D1 = D2 = D3 versus D1 < D2 < D3.

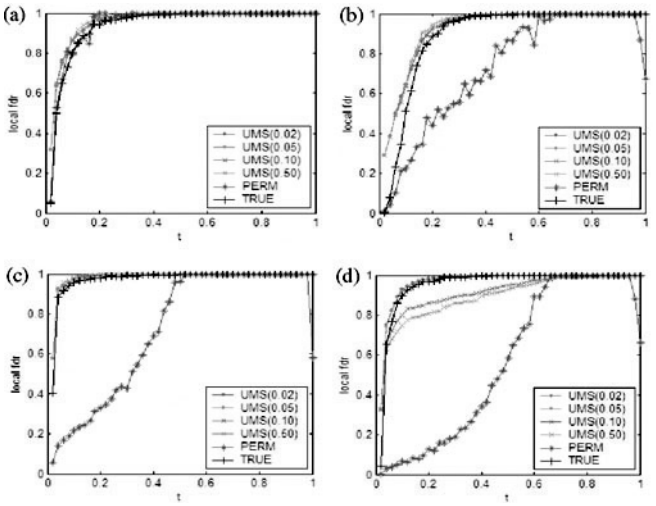

Fig. 4. Local false discovery rate estimates by UMS and permutation test in simulations I–III. $t_{(p)} = t_{(1)}$. UMS was applied under four different $t_{(q)}$ settings: $t_{(q)}$ = 2nd, 5th, 10th, 50th percentile of t. Black curves correspond to true lfdr. (**a**) Simulation I, estimations based on t without variance shrinking; (**b**) simulation I, estimations based on t with variance shrinking; (**c**) simulation II, estimations based on t without variance shrinking; (**d**) simulation III, estimations based on t with variance shrinking.

Probe level test-statistics t were computed for three-sample comparison D1 < D2 < D3. UMS was applied to estimate π_1 and lfdr based on t. In UMS, $t_{(p)} = t_{(1)}$, $t_{(q)}$was set to $t_{(2)}$, $t_{(5)}$, $t_{(10)}$ and $t_{(50)}$ respectively, and [0, 1] was divided into $n = 50$ intervals. For comparison's purpose, permutation test was also applied to estimate lfdr. Since we knew exactly what probes were true targets, the true lfdr could be obtained. Both true lfdr and lfdr obtained by permutation test were shown together with UMS estimates in Figure 4a and b. In Figure 4a, estimations were based on t without variance shrinking. As expected, both UMS and permutation test gave desired lfdr, with UMS being a little bit more conservative. In Figure 4b, estimations were based on t with variance shrinking. Surprisingly, permutation test failed to provide desired lfdr, even though the null hypothesis here was D1 = D2 = D3. This was owing to the combined effect of permutation test and shrinking. The sample variance of probes with D1 < D2 < D3 tend to become bigger after permutations; therefore, variance estimates of all probes were pulled toward a bigger $\overline{s^2}$ when shrinkage estimator was applied, and test-statistics tend to become more centralized in the permutation distribution. As a result, the

number of false probes was underestimated on the tail part, causing optimistic lfdr estimates. In contrast to permutation test, however, UMS still provided conservative lfdr estimates.

In simulation II (Section 4.2, Supplementary material), probes outside target regions were assigned some random changes. This introduced random components, such as D1 $<$ D2 $>$ D3, D1 $>$ D2 $<$ D3 into the null hypothesis which is no longer D1 $=$ D2 $=$ D3. UMS and permutation test were both applied to estimate lfdr, and the results based on t without and with variance shrinking were shown in Figure 4c and Supplementary Figure S3b , respectively. Now even in the non-shrinking case, permutation test failed to provide desired lfdr for D1 $<$ D2 $<$ D3. UMS, however, again provided conservative estimates for both non-shrinking and shrinking t under different $t_{(q)}$ settings.

Simulations I and II tested UMS when its conditional independence assumption [i.e. $f(T_i = t|H_i, I_{\{u_i \in A\}} = 1) = f(T_i = t|H_i)$] was true. Analysis of Cawley's experiment showed that this assumption can provide a first order approximation of the real data (Section 5.1, Supplementary material). To see how UMS performs when this assumption does not hold, in simulations III–VI, we challenged UMS by violating the assumption in different ways.

In simulation III (Section 4.3, Supplementary material), we introduced some additional binding regions with pattern D1 $=$ D2 $<$ D3 or D1 $<$ D2 $=$ D3 into the background. Each type of the new regions also covered π_1 of the total probes. The additional regions belonged to null hypothesis and were not the targets we wish to detect. This design broke the conditional independence assumption under $H_i = 0$, since D1 $=$ D2 $<$ D3 and D1 $<$ D2 $=$ D3 are more likely to generate significant test-statistics than D1 $=$ D2 $=$ D3 does, and unlike simulation II, here probes from D1 $=$ D2 $<$ D3 and D1 $<$ D2 $=$ D3 tend to be clustered together. The lfdr estimates by UMS and permutation test in this simulation are shown in Figure 4d and Supplementary Figure S3c. When $t_{(q)}$ was small ($q\% \leq \pi_1$ in this case), UMS was still able to provide reasonable lfdr estimates. When $t_{(q)}$ became large, the estimates became optimistic. In both cases, however, UMS performed much better than permutation test. A theoretical analysis of why UMS works in such a situation when $t_{(q)}$ is small is given by supplementary material (Section 5.2).

Simulations IV–VI (Section 4.4, Supplementary material) were tailored from simulations I–III respectively. Residual correlations between T_i and u_i were introduced into binding regions, which broke the conditional independence assumption under $H_i = 1$. The results obtained were shown in Supplementary Figure S3 and were similar to those in Figure 4, suggesting that this type of violation of the assumption did not influence the performance of UMS significantly.

We did additional theoretical analysis and tests (Sections 4.5–4.8 and Section 5, Supplementary material). Together with simulations here, they showed that (1) when the conditional independence assumption of UMS holds, UMS can provide reasonable lfdr and π_1 estimates, and the performance is robust to choices of $t_{(p)}$ and $t_{(q)}$; (2) when the assumption does not hold, UMS can provide reasonable lfdr and π_1 estimates when $t_{(q)}$ is small, and under such condition, the performance of UMS is robust to choices of $t_{(p)}$; however, if $t_{(q)}$ is big, UMS is sensitive to choices of $t_{(p)}$; (3) UMS works reasonably well when m_i instead of t_i is used to estimate lfdr. According to our own experience, by setting $t_{(q)} \leq t_{(5)}$ and $t_{(1)} \leq t_{(p)} \leq t_{(20)}$, UMS usually can provide reasonable performance.

Finally, when applying UMS to Cawley's experiment, at lfdr $= 0.5$ level, MA detected 30 and 19 regions with pattern IP $>$ C1 and IP $>$ C2 for cMyc (ChipA) and p53-FL data, respectively. At posterior probability $= 0.5$ level, HMM detected 168 and 142 regions. As a comparison, at P-value $= 0.001$ level, G-TRANS reported 48 and 152 regions. HMM tend to report more regions than MA, many of which are regions shorter than the window size specified by MA and are not reported by MA (Section 6, Supplementary material). Whether the shorter regions found by HMM are more likely to be true signals or noise cannot be clearly resolved using current data. When we checked the probe intensities, many such regions did look like true binding regions (Figure S8). Future experimental verifications are needed to resolve this issue.

5 DISCUSSION

Compared with previous tools, TileMap provides a flexible way to study tiling array hybridizations under multiple experimental conditions. The variance shrinking component increases the sensitivity in finding genomic loci of interest when the number of replicates is small. Though we have only illustrated the use of TileMap in ChIP-chip experiment, it can also be used to analyze transcriptional activities of the genome. In terms of computation time, TileMap is substantially more efficient than G-TRANS.

The main difficulty of multiple sample comparisons is to get the distribution of test-statistics under the null hypothesis, which is needed for FDR control or HMM decoding. TileMap adopts an approximate procedure, UMS, to deal with this issue. UMS is not a perfect solution. However, the estimation of null distributions under complex composite null is a difficult problem in general, for which there are no good solutions currently. UMS embodies an initial try to address this issue. The rough estimate provided by UMS can be used to guide the choice of cutoffs, and in many cases, such an imprecise estimate is enough for practical use for several reasons. First, FDR is always model dependent, e.g. assuming $H_0 : \mu_1 = \mu_2$ and $H_0 : \mu_1 - \mu_2 \sim N(0, 1)$ will result in very different FDR estimates. Therefore, unless the statistical null model (e.g. $\mu_1 = \mu_2$) matches the scientific null (e.g. not tumor-related), FDR could be very misleading. Second, compared with power, FDR is of secondary importance if we are only interested in a few top regions. What we really care about is to have higher chance to find regions of real scientific interest instead of getting an FDR estimate for a statistical model which may be an oversimplification of the real world. Despite all these arguments, we also acknowledge that further investigation of how to control FDR under composite null *per se* deserves further investigation. Such investigations will provide basis for rigorous statistical inference for complex multiple sample comparisons.

Both MA and HMM used here did not consider the real distributions of the length of hybridizations. Current knowledge about such distributions is limited. If one can determine these distributions, the models here can be refined and may provide further resolving power. For MA, the average can be replaced by a weighted average; for HMM, a distance dependent transition probability can be used. All these aspects deserve further investigation. Finally, TileMap is only the first step to utilize the information provided by tiling array. Future efforts to integrate TileMap with *cis*-regulatory module discovery, alternative splicing analysis, etc. will help us get deeper understanding of various biological systems.

ACKNOWLEDGEMENTS

The authors thank Simon Cawley for providing the chromosome 21 and 22 ChIP-chip data, Xiaole S. Liu for helpful discussions about using HMM in analyzing tiling arrays, and the two anonymous referees for their invaluable suggestions to improve the paper. The work was partially supported by NIH grant GM-067250. Funding to pay the Open Access publication charges for this article was provided by the same grant.

Conflict of Interest: none declared.

REFERENCES

Baldi,P. and Long,A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509–519.

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.

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.

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.

Kapranov,P. *et al*. (2003) Beyond expression profiling: next generation uses of high density oligonucleotide arrays. *Brief. Funct. Genomic. Proteomic.*, **2**, 47–56.

Keles,S., van der Laan,M.J., Dudoit,S. and Cawley,S.E. (2004) Multiple testing methods for ChIP-Chip high density oligonucleotide array data. *Working Paper Series, Paper 147*, U.C. Berkeley Division of Biostatistics, University of California, Berkeley, CA.

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.

Morris,C.N. (1983) Natural exponential families with quadratic variance functions: statistical theory. *The Annals of Statistics*, **11**, 515–529.

Newton,M.A. *et al*. (2004) Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, **5**, 155–176.

Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article 3.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Tusher,V.G. *et al*. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.