Genome analysis

Double error shrinkage method for identifying protein binding sites observed by tiling arrays with limited replication

Youngchul Kim^{1,2}, Stefan Bekiranov³, Jae K. Lee^{1,*} and Taesung Park^{2,*}¹Department of Public Health Sciences, University of Virginia, Charlottesville, VA 22908, USA, ²Department of Statistics, Seoul National University, Seoul, Korea and ³Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA 22908, USA

Received on April 15, 2009; Revised on July 21, 2009; Accepted July 30, 2009

Advance Access publication August 10, 2009

Associate Editor: Alex Bateman

ABSTRACT

Motivation: ChIP–chip has been widely used for various genome-wide biological investigations. Given the small number of replicates (typically two to three) per biological sample, methods of analysis that control the variance are desirable but in short supply. We propose a double error shrinkage (DES) method by using moving average statistics based on local-pooled error estimates which effectively control both heterogeneous error variances and correlation structures of an extremely large number of individual probes on tiling arrays.

Results: Applying DES to ChIP–chip tiling array study for discovering genome-wide protein-binding sites, we identified 8400 target regions that include highly likely TFIIID binding sites. About 33% of these were well matched with the known transcription starting sites on the DBTSS library, while many other newly identified sites have a high chance to be real binding sites based on a high positive predictive value of DES. We also showed the superior performance of DES compared with other commonly used methods for detecting actual protein binding sites.

Contact: tspark@snu.ac.kr; jaeklee@virginia.edu

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

1 INTRODUCTION

The chromatin immunoprecipitation on chip (ChIP–chip) technique, for obtaining information on which transcription factors directly regulate specific genes, has been widely used to discover protein-binding sites on DNA sequences. These binding sites may indicate functions of various transcriptional regulators and help in identifying their target genes. Dysregulation of a transcription factor (e.g. cMyc) can be a cause for human diseases including cancer. Newly-identified binding sites may also provide additional annotation and functional elements in genomes, such as promoters, enhancers, repressor and silencing elements, insulators, boundary elements, and sequences that control DNA replication. ChIP–chip was first successfully adopted in yeast to identify the regulatory targets of individual transcription factor binding sites (TFBS) (Lieb *et al.*, 2001; Ren *et al.*, 2000) and to study entire transcriptional regulatory networks (Harbison *et al.*, 2004). More recently, a high-density

tiling array-based ChIP–chip technique has been utilized for a whole genome-wide localization of transcription factor binding sites, origins of replication, DNase hypersensitive site, DNA methylation, and histone mappings (Cawley *et al.*, 2004; Kim *et al.*, 2005; Sun *et al.*, 2003). The ChIP–chip approach has also been extensively used by the ENCODE consortium to map functional elements in the human genome (<http://www.genome.gov>).

Several methods have been proposed for analyzing ChIP–chip data. **ChIPOTle** (Buck *et al.*, 2005) is based on a simple moving average statistic on sliding windows, considering highly variable characteristics of tiling array data. Assuming log ratios of intensities are i.i.d. from a standard Gaussian distribution, it can also provide corrected *p*-values of candidate sites for multiple comparisons. **Mpeak** (Kim *et al.*, 2005; Zheng *et al.*, 2007) is based on a likelihood ratio test for detecting protein binding region based on double regression model with two symmetric slopes, assuming independence of expression intensities among different probes. **ChIPmix** (Martin-Magniette *et al.*, 2008) is based on a linear regression mixture model, directly using the original signals of each probe for a target ChIP sample conditional on that of a control sample. This approach, however, is reported to occasionally fail to fit a good regression model due to inappropriate initial parameter settings and takes considerable computing time due to the slow convergence of its EM algorithm. **MA2C** (Song *et al.*, 2007) is a model-based approach to analyzing two-color tiling microarray data, incorporating sequence-specific probe effects and peak detection algorithms. It provides a novel normalization method based on the GC content of probes on genome sequence. Also, it detects the protein-binding peaks using a window-based approach, proposed by Johnson *et al.* (2006), which is based on the distribution of polished median and trimmed mean in the window. In the peak detection step, MA2C assumes that its final scores, MA2Cscores, approximately follow a normal distribution with the representative scores of peak regions corresponding to the right tail. Finally, **CMARRT** (Kuan *et al.*, 2008) was proposed to extend ChIPOTle to consider correlations between adjacent probes using autocorrelation coefficients. In this study, the authors used constant variance estimates based on background or noise intensities from control sample in order to avoid reflecting highly variable ChIP-specific effects on error estimates.

Even though these previous methods have attempted to overcome the difficulties of analyzing extremely variable tiling array data and

*To whom correspondence should be addressed.

ChIP–chip experiment, none has reflected the heterogeneous error variance and correlation structures for numerous tiling array probes. Without understanding and using such heterogeneous variance information, a statistical discovery method may inevitably result in a high false positive rate and a low statistical power for detecting real regulatory binding sites. It is important to take the spatial correlation structure into account due to the high density of probes in currently available tiling arrays (Martin-Magniette *et al.*, 2008).

Thus, we attempt to fully utilize the error variance and correlation structures of ChIP–chip tiling array data and propose a double error shrinkage (DES) approach based both on local pooled error (LPE) estimates and correlated moving average statistics. The LPE estimation procedure addresses the heterogeneous errors of numerous tiling array probes with statistical rigor and fidelity, especially in studies with a small number of replicates (Jain *et al.*, 2003). The moving window step is also useful for discovering real protein binding sites considering their spatially correlated characteristics. We compare the performance of DES with other commonly-used methods for detecting protein binding sites in a recent large discovery study of 38 ChIP–chip tiling array data on fibroblast IMR90 cells (Kim *et al.*, 2005).

2 METHODS

2.1 Local pooled error for ChIP–chip array data

Individual probe intensities in tiling arrays are highly variable. Therefore, it is important to consider the initial analysis step that can evaluate individual probe intensities more reliably by reducing noise levels. In order to derive accurate statistical estimation of heterogeneous errors of individual probes in tiling arrays, we adopt a recent error pooling technique—local pooled error (LPE)—which pools the error information in each local intensity range and shrinks each error variance estimate toward the mean of hundreds of other probes with similar intensity levels (Jain *et al.*, 2003; Park *et al.*, 2007). However, the original LPE relies on replicated microarray experiments and cannot be directly applied to ChIP–chip tiling arrays, especially when no replicated array data are available. Thus, we modify the LPE method as described below.

We describe here our LPE extension for the two-channel tiling arrays (e.g. NimbleGen), but this will essentially be identical for the single-channel tiling arrays, e.g. in Affymetrix tiling arrays using experimental and control sample arrays.

Let R and G represent the intensities of each probe from the experimental (e.g. with Cy5) and control (with Cy3) samples on two-color tiling arrays, respectively. The paired intensity values (R , G) of each probe are then converted into a log ratio as follows:

$$M = \log_2 \left(\frac{R}{G} \right), \quad A = \frac{(\log_2 R + \log_2 G)}{2}$$

M is thus the relative abundance (or difference) of each probe and A the average intensity between the experimental and control samples. In order to minimize instrumental error variation of individual probes, we use the Lowess normalization method based on the locally weighted estimation function of A versus M (Yang *et al.*, 2002). Note that the M values are expected to be highly positive if probe sequences contain actual protein-binding sites; otherwise, such values will be distributed as random noise distributions. Thus, almost all negative M values are from random noise since such a negative value would not likely be observed if there is an actual binding site and the corresponding tiling array probe can detect such a site. Therefore, the null error distribution can be well derived using these negative M values (M_-) and their mirror reflection. In order to reduce the effects of outliers, we also use a robust outlier detection procedure commonly used to detect outliers in Tukey's Box-plot, removing extreme outliers based on each

probe's scale parameters such as global standard deviation and inter-quartile range indices (Li *et al.*, 2005). We then estimate the modified LPE variances. Briefly, the negative M values within

$$(Q_{1M_-} - 3 \times IQR_{M_-}, Q_{3M_-} + 3 \times IQR_{M_-})$$

are used to estimate LPE variance; here, Q_{1M} and Q_{3M} are the first and third quartiles of negative M values at each local intensity region and the interquartile range (IQR) is the difference between these two quartiles. Note that we use all M_- values in each fixed-width window intensity interval of A based on pre-specified percentiles to compute the error variance estimates as the following:

$$\hat{\sigma}_i^2 = (n_i - 0.5)^{-1} \sum_{j=1}^{n_i} M_{ij-}^2$$

where M_{ij-} is the j -th M_- value in the i th quantile interval of A and n_i is the total number of M_- values in the i -th interval of A . A nonparametric smoothing function is then fitted with the LPE estimated variances across the entire intensity range. For generalizing the analysis to the single channel tiling arrays like Affymetrix microarray platform, we define (M, A) with the two perfect match intensities from normalized experimental and control sample arrays instead of R and G of the two-channel tiling array. The above modified LPE variance estimation can then be directly applied to these (M, A) values (Park *et al.*, 2007).

2.2 Weighted moving average statistics for robust intensity estimation of binding sites

Since the immunoprecipitated DNA sequences were irregularly sheared in ChIP samples, these sequences would be randomly hybridized to the adjacent sequence locations of each protein-binding site in a symmetric unimodal fashion. In order to utilize such a unimodal distribution and dependency among adjacent probes in identifying binding sites, we use an (auto-covariance) weighted moving average (WMA) statistic within each fixed-length sliding window. This WMA statistic is defined based on LPE variance estimates and their auto-covariance estimates as follows:

$$T_i = \frac{S_i}{\sqrt{\text{Var}(S_i)}},$$

where,

$$S_i = \frac{\sum_{j=1}^K w_{ij} M_{ij}}{\sum_{j=1}^K w_{ij}}, \quad \text{Var}(S_i) = \frac{\sum_{j=1}^K w_{ij}^2 \sigma_{ij}^2}{\left(\sum_{j=1}^K w_{ij} \right)^2} + 2 \sum_{j=1}^{K-1} \sum_{k=j+1}^K \frac{w_{ij} w_{ik}}{\left(\sum_{j=1}^K w_{ij} \right)^2} \sigma_{ijk}.$$

K is the number of probes in the sliding window and w_{ij} and M_{ij} are the intensity and weight for the j -th probe within the i -th sliding window, respectively. The symmetric kernel function around the center of the moving window such as uniform, quadratic, exponential, and Gaussian kernels were used to compute the weight of each probe by putting the standardized genomic distance between the corresponding probe and the center position within the sliding window. σ_{ij}^2 is the variance of the j -th probe intensity within the i -th sliding window and σ_{ijk} the auto-covariance of the j -th and the k -th probe intensities within the i -th sliding window. When the WMA statistic is calculated, LPE variance estimates for the j -th probe are used, just as the estimates of σ_{ij}^2 and auto-sample covariance estimates are also used as the estimates of σ_{ijk} .

This WMA statistic thus combines information of adjacent probes and different degrees of probe-specific effects within a given window. Experimenting with various window sizes 100 to 1kbp and weights (by balancing positive predictive value and statistical power; data not shown), we chose the constant weights and the optimal window size 500 bp (or about five consecutive tiling array probes on average). Note that this statistic can also account for heterogeneous error variances of different probe sequences by averaging the different LPE variance estimates of individual probes in each window. We can rapidly get a p -value approximation for each window

by a parametric estimation under the assumption that log ratios are random variables having a Gaussian distribution with a mean of zero as follows.

$$P_i = 1 - \Phi(T_i),$$

where Φ is the cumulative distribution function of standard Gaussian distribution. We then use a corrected p -value for multiple comparisons using FDR adjustment. Note that when the sliding windows overlap across consecutive probes, the dependency of the p -values may skew the FDR estimate. The permutation-based non-parametric test is then performed on these WMA statistics to control the false discovery rate (FDR) for selecting significant moving window locations of candidate protein-binding sites (Storey and Tibshirani, 2003).

3 RESULTS

3.1 ChIP-chip data on IMR90 cells

A ChIP-chip study using high-density tiling arrays was performed to identify genome-wide transcription binding sites on the fibroblast IMR90 cells by Kim *et al.* (2005). Briefly, the 38 arrays covered the whole human genome with a total of 14 535 659 50-mer oligonucleotides at every 100 bp apart. To identify active promoters in human cells, DNA bound by the general transcription factor IID (TFIID) from cross-linked primary fibroblast IMR90 cells was isolated by chromatin immunoprecipitation with an antibody which specifically recognizes the TAF1 subunit of this complex (TBP associated factor 1, formerly TAF250). The enriched DNA was then amplified and fluorescently labeled and hybridized to the high-density oligonucleotide arrays along with a differentially labeled control DNA. The hybridized arrays were scanned on an Axon GenePix 4000B scanner (Axon Instruments Inc.) at wave lengths of 532 nm for control (Cy3), and 635nm (Cy5) for experimental sample. Data were extracted from the scanned images using the NimbleScan 2.0 program (NimbleGen Systems, Inc.). The microarray data sets are available from GEO database (Gene Expression Omnibus; GSE2672) and from <http://licrrenlab.ucsd.edu/download.html>.

3.2 LPE estimation for tiling array probe intensities

The NimbleGen tiling array used in the above study was designed with about 6000 control probes (1.3%) out of 387 000 probes. These probes showed very little intensity changes and were first removed in our analysis. The correlation coefficients between the control sample (Cy3) and the ChIP sample (Cy5) across all tiling arrays varied from 0.84 to 0.98, implying a high reproducibility of this ChIP-chip experiment. Based on the outlier detection method described earlier, about 5106 probes (2.6%) were classified as outliers and were removed prior to estimating the LPE variance functions.

Figure 1a shows various M versus A plots of individual arrays. Figure 1b shows LPE variance functions for all 38 ChIP-chip tiling arrays. It can be easily seen that the assumption of homoscedasticity of error variance for all probes was not appropriate for these tiling arrays. The LPE functions showed several distinctive patterns (Fig. 1c); the first representative pattern was a monotone decrease pattern, the second an up-down pattern, and another a down-up-down pattern. These effects seem to be strongly array-dependent. This is not a surprise given that these arrays interrogate local chromosomal regions with different GC content.

Fig. 1. LPE variance function from ChIP-chip without replication. (a) MA scatter plot. (b) LPE variance functions for total 38 ChIP-chips. (c) Three distinctive patterns of LPE variance functions for single ChIP-chip arrays.

3.3 Detection of significant probes through weighted moving average test statistic

As described in the ‘Methods’ section, performing the permutation test by shuffling the probe intensities within each tiling array, we obtained empirical p -values, and then q -values to control the false discovery rate (FDR) (Storey and Tibshirani, 2003). Figure 2 is a Q-Q plot for Gaussian distribution versus a TFIID ChIP-chip. It shows that Gaussian approximation approach could also be useful to obtain the p -value. Also, those TFIID binding sites identified in Kim *et al.* (2005) were validated through DBTSS (database of transcription start sites, Suzuki *et al.*, 2004) and are used as true positives to assess the performance of DES.

We experimented with different sliding window sizes and weight types such as the uniform, quadratic, triangular, Gaussian, and Epanechnikov kernels. Using the window size 500 bp and constant weight, we found that DES provided an optimal performance with a low false positive rate and high power of detecting protein binding sites for the IMR90 data (see Supplementary Tables S1 to S3). It resulted in 96 745 total significant probes with $FDR < 0.1$ which were further narrowed to 8400 candidate binding regions by combining consecutively identified candidate binding probe sites.

The (auto-correlation) weighted moving-average step of DES has the effect of greatly reducing spurious noise near the binding sites compared with the raw profiles of M . We found that significant

Fig. 2. Q-Q plot for TFIID ChIP-chip versus Gaussian distribution.

probes from this step tended to cluster together since many sequences of various sizes exist which bind the target protein. We note that the profiles of the moving average show either a unimodal or bimodal pattern around these regions. Also the size of clusters varied significantly across the genome location even though the size of significant probe clusters at protein-binding sites were expected to be equal or similar, since this ChIP-chip tiling array study was performed for IMR90 cells only with TFIID protein-binding sites. There were some clusters that included only one significant probe located far from others; these clusters generally showed low-intensity peaks and are likely artifacts and not real binding sites.

3.4 Comparison of performance and computing time

We examined the performance of our DES and other commonly used methods, ChIPOTle, ChIPmix, CMARRT and MA2C, for analyzing tiling ChIP-chip data under the same FDR level and sliding window size. We then compared our results with others based on positive predictive value (PPV) and sensitivity. We note that other measures of performance such as specificity and negative predictive value (NPV) were not informative for our comparison since almost all tiling array probes are from non-binding sites and an extremely small proportion of probes are binding sites; in fact, specificity and NPV were found to be almost near 1 for all the above methods when applied to the IMR90 cell tiling array data. Also note that sensitivity here was computed based only on 10 530 true transcription start sites found in the DBTSS library which is not comprehensive and does not include all true binding sites. On the other hand, after fixing the level of sensitivity, PPV can be an objective measure to compare their discovery performance since this measure can show the enriched proportion of true binding sites among all candidate sites identified by each of these methods when the same level of statistical power is sought by them.

Also, as done with our DES approach above, consecutively identified probes by each of CMARRT, ChIPOTle, ChIPmix and MA2C were combined as a single binding site for our comparison. This strategy was found to greatly reduce the number of false positives in the above studies and our experiments. Another important factor in our comparison was to remove sporadic candidate sites with only a small number of consecutively identified

Fig. 3. PPV versus minimum consecutively identified probes.

Fig. 4. Comparison of PPV at varying sensitivity levels.

probes. Therefore, we investigated the performance of the above methods uniformly applying the same cutoff criterion of removing the candidate sites identified with less than a certain number (2~30) of consecutive probes by each method (Fig. 4). If a candidate site (from its center location) by each method was located within 250 bp from the nearest transcription start site in DBTSS, that site was counted as a correctly detected binding site for our PPV and sensitivity evaluation.

Figure 3 shows the trend of PPVs of the five methods when the candidate sites were filtered only with the ones identified by more than a certain number of consecutive probes (size). While CMARRT and ChIPOTle, which are also based on moving average statistics, show similar performance as DES, DES significantly outperformed other methods. (Two sample binomial p -values $< 1e-4$). As the size of consecutive regions increases, PPV also increases up to 0.5 but decreases when the size of those regions exceeds 17. It means that the candidate binding regions having too many probes not displaying the hallmarks of a ChIP enrichment profile are likely artifacts. Consequently, DES provided relatively shorter and more accurate candidate protein binding regions compared to the other methods.

Figure 4 shows the PPVs at varying sensitivity levels of the five methods by varying the size cutoff values above. DES again had the best PPV at all sensitivity levels among the five with the highest PPV over 40% compared with 34% or lower of the others.

Table 1. Performance and computing time

Method	Number of correctly detected/total number of candidate regions	PPV (%)	Sensitivity (%)	Computing time (s)	Soft ware
DES	2750/8400	32.73	26.12	8	R
CMARRT	2663/8436	31.56	25.28	37	R
ChIPOTle	2705/8685	31.14	25.69	476	Perl
Mpeak	2595/9030	28.74	24.64	4	C
ChIPmix	2725/57 422	4.74	25.87	> 1800	R
MA2C	2600/491 030	0.53	24.69	56	JAVA

Table 1 summarizes the performance indices and computing times of the five methods under the same FDR level 10%, a fixed cutoff size of 10 consecutive probes (about 1 kb), and the same level of sensitivity. At a fixed sensitivity level 26%, DES identified 2750 real binding regions out of 8400 candidates with its PPV 32.7%. CMARRT provided 2663 correctly detected binding regions out of 8436 candidates with PPV 31.6% at sensitivity 25.3%. ChIPOTle also resulted in 2705 correctly detected binding regions out of 8685 candidates with PPV 31.1%. ChIPmix showed a very low PPV 4.7% at cutoff value 9 in which it showed a similar sensitivity level with others. MA2C showed the lowest PPV 0.5% and sensitivity 24.7% because the distribution of MA2Cscores on some of ChIP-chip is left-skewed. The low performance of MA2C may be due to the fact that this algorithm has not been optimized for one-color ChIP-chip data structure. While Mpeak may not need to be adjusted by the size cutoff criterion because it returns a single probe as each candidate, it identified 9030 candidate sites, 2595 of which were correctly detected with a PPV of 22%. These results thus showed that DES would identify candidate binding regions with the highest accuracy among the compared methods.

In practical applications, the computing time and resources are often important when this kind of search is performed repeatedly for a large data set. Reiss *et al.* (2008) proposed MeDiChI which models the single protein binding event and finds the peak profile. However it needs the deconvolution of each ChIP-chip data set into smaller segments and the computing cluster or multiple processor cores to apply high-throughput tiling ChIP-chip.

DES, CMARRT, and ChIPmix were all implemented in R (a statistical programming language/environment), whereas Mpeak, MA2C and ChIPOTle were implemented with Perl, JAVA and C codes, respectively. DES took about 8 s to finish analyzing a single ChIP-chip array compared to 37 s by CMARRT and >30 min by ChIPmix on the same data. ChIPmix sometimes failed to converge for the final model. Mpeak implemented in Microsoft windows application provided the fastest computing time.

3.5 Comparison of performance in a simulated ChIP-chip experiment

Johnson *et al.* (2008) provided very valuable ChIP-chip data based on well simulated ChIP spike-ins mixture. They then created various enrichment levels that 1.25-fold to 196-fold relative genomic DNA. This data set is available from GEO database (Gene Expression Omnibus; GSE10076). Again, we examined the performance of DES and other methods, ChIPOTle, ChIPmix, CMARRT and Mpeak at

Fig. 5. PPV versus minimum consecutively identified probes. **Fig. 6.** Comparison of PPV at varying sensitivity levels.

FDR level 0.05 except for MA2C due to the lack of probe sequence information.

Figure 5 shows the trend of PPVs of the four methods when the candidate sites were filtered only with the ones identified by more than a certain number of consecutive probes. It is similar with the result on TFIID ChIP-chip. DES also provided relatively short and more accurate candidate spike-ins locations compared to other methods.

Figure 6 shows the PPVs at varying sensitivity levels of the four methods by varying the size cutoff values. Mpeak showed the PPV 25.49% and sensitivity 65%. DES again had the best PPV at all sensitivity levels among the four with the highest PPV over 60% at a fixed sensitivity level about 65%.

4 DISCUSSION

In this study, we introduced the DES method based on local pooled error estimates and auto-correlation weighted moving average statistics to discover protein binding sites from ChIP-chip tiling data without replication. DES showed a significant improvement in identifying real binding sites compared to other alternative methods. It is worth noting that the statistical power and PPV of DES were equivalent to or higher than those obtained by Kim *et al.* (2005) who performed not only this ChIP-chip experiment but also an expensive

follow-up customized microarray experiment with initial candidate 10K probes from their ChIP–chip data analysis.

The major strengths of DES are its accounting of heterogeneous error variances of numerous tiling array probe intensities, in contrast to a constant variance assumption used in other approaches, as well as correlations of adjacent tiling array probes, especially near true binding sites. These error shrinkage techniques minimize noise in ChIP–chip data as well as false discovery rates due to the extremely high number of multiple comparisons in tiling array data analysis.

Both the sliding window and the weighted average methods played important roles of reducing the effect of specific erroneous probes and hence in estimating the distribution of intensities near protein-binding sites more accurately. In order to see the effects of these parameters, we investigated the performance of the proposed DES method with various sliding window sizes and weight types. In summary, the shorter the window size, the higher the sensitivity but the lower the PPV. However, the weight type did not show a large effect on the result.

It has been observed by us and others that selecting the candidate sites based on the number of consecutively significantly identified probes is a very effective filtering step. Although this filtering step can be an important issue in the analysis of tiling arrays, no formal statistical method has been considered with this criterion and its future study is warranted.

Our DES method can be applied to various kinds of ChIP–chip studies in addition to TFIID study. For example, it could be used in the genome-wide identification of histone modifications. Note that the original ChIP–chip study was developed to locate promoters, enhancers and insulators in the human genome. Especially, if ChIP–chip is used to identify enhancers in multiple cell types and in investigating their roles in cell-type-specific gene expression, it helps to identify cell-type-specific histone modification patterns (Heintzman *et al.*, 2009).

Our DES was firstly developed for the analysis of single ChIP–chip experiment. However, the principal components of DES, WMA statistics and LPE variance estimator could be applicable to replicated experiment as described in Jain *et al.* (2003) and Park *et al.* (2007). There are several ways of applying DES to the replicated experiments. The first simple one is to take average of replication and then apply DES. The second one is to apply DES for each replication by using the LPE error variance based on all pair-wise negative M values, and computing WMA test statistics for each replication. Then, combine the results, say by taking the intersection regions of the identified protein-binding regions. In order to see the performance on these cases, we need to develop more a rigorous approach based on the moving averages across all of the replicates.

We believe that this study demonstrates that an improved analysis strategy could significantly increase statistical power for identifying real binding sites from tiling array data with a limited replication. However, various analysis steps in DES may require continuous refinement by applying and investigating its performance in a broader range of applications.

ACKNOWLEDGEMENTS

The authors thank three anonymous reviewers for their helpful comments. The authors also thank Kyung Kuk Choi for his part in editing.

Funding: US NIH grant 5R01HL081690 to J.K.L.; National Research Laboratory Program of Korea Science and Engineering Foundation (M10500000126) to T.P.; Next-Century Korea Scholarship Program of Korea Research Foundation (C100015) to Y.C.K.

Conflict of Interest: none declared.

REFERENCES

- Buck, M.J. *et al.* (2005) ChIPoTle: a user-friendly tool for the analysis of ChIP–chip data. *Genome Biol.*, **6**, R97.
- 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.
- Harbison, C. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
- Heintzman, N.D. *et al.* (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**, 108–112.
- Jain, N. *et al.* (2003) Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics*, **19**, 1945–1951.
- Johnson, D. *et al.* (2008) Systematic evaluation of variability in ChIP–chip experiments using predefined DNA target. *Genome Res.*, **18**, 393–403.
- Johnson, W. *et al.* (2006) Model-based analysis of tiling-arrays for ChIP–chip. *Proc. Natl Acad. Sci.*, **103**, 12457–12462.
- Kim, T.H. *et al.* (2005) A high-resolution map of active promoters in the human genome. *Nature*, **436**, 876–880.
- Kuan, P. *et al.* (2008) CMARRT: A tool for the analysis of ChIP–chip data from tiling arrays by incorporating the correlation structure. *Pac. Symp. Biocomput.*, **13**, 515–526.
- 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.
- Lieb, J.D. *et al.* (2001) Promoter-specific binding of Rap1 revealed by genome-wide maps of protein–DNA association. *Nat. Genet.*, **28**, 303–304.
- Martin-Magniette, M.L. *et al.* (2008) ChIPmix: mixture model of regression for two color ChIP–chip analysis. *Bioinformatics*, **24**, i181–i186.
- Park, T.S. *et al.* (2007) Error pooling-based statistical methods for identifying novel temporal replication profiles of human chromosomes observed by DNA tiling arrays. *Nucleic Acids Res.*, **35**, e69.
- Reiss, D.J. *et al.* (2008) Model-based deconvolution of genome-wide DNA binding. *Bioinformatics*, **24**, 396–403.
- Ren, B. *et al.* (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.
- Song, J.S. *et al.* (2007) Model-based analysis of two-color arrays (MA2C). *Genome Biol.*, **8**, R178.
- Storey, J.D. and Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.
- Sun, L.V. *et al.* (2003) Protein–DNA interaction mapping using genomic tiling path microarrays in *Drosophila*. *Proc. Natl Acad. Sci.*, **100**, 9428–9433.
- Suzuki, Y. *et al.* (2004) DBTSS, Database of transcriptional start sites: progress report. *Nucleic Acids Res.*, **32**, D78–D81.
- Yang, Y.H. *et al.* (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.
- Zheng, M. *et al.* (2007) ChIP–chip: data, model, and analysis. *Biometrics*, **63**, 787–796.