Gene expression

Meta-analysis based on control of false discovery rate: combining yeast ChIP-chip datasets

Saumyadipta Pyne\textsuperscript{1,*}, Bruce Futcher\textsuperscript{2} and Steve Skiena\textsuperscript{1}
\textsuperscript{1}Department of Computer Science and \textsuperscript{2}Department of Molecular Genetics and Microbiology, Stony Brook University, NY 11794, USA

Received on June 15, 2006; revised on August 2, 2006; accepted on August 9, 2006
Advance Access publication August 14, 2006
Associate Editor: Martin Bishop

ABSTRACT

Motivation: High-throughput microarray technology can be used to examine thousands of features, such as all the genes of an organism, and measure their expression. Two important issues of microarray bioinformatics are first, how to combine the significance values for each feature across experiments with high statistical power, and second, how to control the proportion of false positives. Existing methods address these issues separately, in spite of their linked usage.

Results: We present a novel method (ESP) to address the two requirements in an interdependent way. It generalizes the truncated product method of Zaykin et al. to combine only those significance values which clear their respective experiment-specific false discovery restrictive thresholds, thus allowing us to control the false discovery rate (FDR) for the final combined result. Further, we introduce several concepts that together offer FDR control, high power, quality control and speed-up in meta-analysis as done by our algorithm. Computational and statistical methods of research synthesis like the one described here will be increasingly important as additional genome-wide datasets accumulate in databases.

We apply our method to combine three well-known ChIP-chip transcription factor binding datasets for budding yeast to identify significant intergenic regulatory sequences for nine cell cycle regulating transcription factors, both with high power and controlled FDR.

Contact: spyne@cs.sunysb.edu
Supplementary Materials and Appendices: http://www.cs.sunysb.edu/~combio/Meta

1 INTRODUCTION

High-throughput microarray technology can be used to examine thousands of features, such as all the genes of an organism, and measure their expression under various experimental conditions. As a result, rapidly growing collections of large datasets are becoming available for subsequent analysis (Moreau et al., 2003). Given the differences in characteristics of the raw datasets, meta-analysis of experiments with high-level data helps not only to identify the consistently true signals but also to annul possible cross-platform effects. The final aim is to identify features of possible biological interest. Hence a major concern of microarray bioinformatics is to combine the significance values for each feature across a collection of independent analogous experiments with high power, while controlling the proportion of false positives among those features declared significant.

Existing methods address these issues independently and sequentially, as e.g. in Rhodes et al. (2002). The variation in the quality of the experiments to be combined is an important concern for genome-wide meta-analysis. Although multiple testing of a genome-wide experiment can help in gauging the overall significance of its results and identify false discoveries, it could be difficult to distinguish many of the combined true positive features after the significance values of each feature over several experiments are coalesced into one omnibus meta-analysis statistic. Therefore we present a method to address the two issues in a co-ordinated way.

The input to our method is an $N \times L$ matrix $M$ such that the $(i,j)$-th entry is a $p$-value for feature $i$ due to experiment $j$. For microarray datasets, typically $N$ runs into several thousands (e.g. all genes), while $L$, the number of experiments to be combined (not to be confused with the number of samples in the microarrays), is a relatively small constant. Our algorithm works as follows: for a specified false discovery level, each column of $p$-values yields an experiment-specific cutoff; these cutoffs are then applied to compute thresholded Fisher product over the row of $p$-values for every feature such that the combined FDR requirement is satisfied by the output set of significant features.

While attempting to integrate results from every experiment, it is difficult to guard against false negatives. Microarray data are often noisy, and the experimental imperfections lead to unreliable as well as missing entries in the data matrix. Even a single sufficiently poor entry, possibly spurious, could skew the combined statistic of an otherwise truly significant feature enough to prevent any test of the joint null hypothesis from rejecting it, thereby forcing the test to lose power. Given the large number ($N$) of features, it can thus lead to many false negatives. To fix this issue, the Fisher product, given by the omnibus multiplication of $p$-values, was extended in Zaykin et al. (2002) using a thresholding criterion whereby only those $p$-values which ‘clear’ (i.e. are less than or equal to) some pre-specified cutoff value $\tau$ contribute to the combined product.

The cutoff ($\tau$) obviously helps to increase the power of the test by letting the poor $p$-values of a feature be ignored. Given its usefulness, it seems natural to extend the choice of $\tau$ such that it is neither arbitrary nor necessarily the same for all the experiments. Indeed, different experiments vary tremendously in quality and in noisiness, and so it is highly desirable to have different $p$-value thresholds associated with different experiments. We suggest a choice of $\tau$ guided by the critical aim of controlling the overall FDR.
Several techniques are known to obtain meaningful p-value cutoffs for genome-wide lists of microarray results, e.g. Storey and Tibshirani (2003). For a chosen FDR level α, we can thus obtain a p-value cutoff τ_{j,α} for each experiment j, where an individual experiment’s FDR α’ is such a value that would bound the combined FDR over L experiments to α. We generalize the thresholded Fisher product in Zaykin et al. (2002) whereby only those p-values which clear their respective experiment-specific, and now likely to be distinctive, cutoffs form the present product (ESP, or experiment-specific product) for which we also compute the probability distribution and hence the combined probability. The features identified as significant as a result of this metaanalysis are guaranteed to satisfy the combined FDR level α.

A global parameter C allows us to impose a consensus requirement that every combined significant feature must have at least C p-values which clear their respective cutoffs. Consensus and thresholding together provide two-pronged control of both false positives and false negatives in the output of our algorithm. Further, Chernoff-Hoeffding probabilistic tail bounds are used for every feature to limit the search for subsets of p-values which clear their cutoffs thereby potentially speeding up the computation of its combined probability. Together, these techniques offer FDR control, high power, quality control and speed-up in meta-analysis.

We apply our algorithm to simulated data as well as to data from three well-known genome-wide transcription factor binding ChIP-chip experiments for budding yeast (Harbison et al., 2004; Lee et al., 2002; Simon et al., 2001). The results of our algorithm were validated by comparing with a fourth dataset (Iyer et al., 2001). Our algorithm yielded better correlation, and presumably a more significant set of output features, than the ordinary Fisher product.

2 COMBINING SIGNIFICANCE VALUES

For any feature g, the corresponding row of M is a vector of L independent p-values \{p_1, p_2, \ldots, p_L\} that represent the tail probabilities due to single hypothesis testing of the experimental results for g. Meta-analysis then helps us obtain a combined significance value to test whether the joint null hypotheses is true. The significance values are often reported in literature in the form of p-values. When the significance is not reported as p-values, the p-values can often be computed with exact or simulated methods.

2.1 Multiplying probabilities

Combining L independent and uniformly distributed p-values \{p_j: j=1,2,..,L\} in order to generate a single probability value has been well studied [surveys include Becker (1994) and Rosenthal (1991)]. Fisher’s inverse \(\chi^2\) method with omnibus product is perhaps the most popular technique. Fisher (1932) observed that under the assumption of uniformly distributed random variable \(p_j \sim U[0,1]\), has a \(\chi^2\) distribution with two degrees of freedom, and hence the statistic \(-2 \ln \prod_{j=1}^{L} p_j\) follows a \(\chi^2\) distribution with 2L degrees of freedom when the joint null is true.

Although omnibus multiplication of probabilities has been the classical approach in meta-analysis to combine results across experiments, not all experiments are created equal. Weighted combination is thus an option, and techniques for doing so have been suggested both for Fisher product in particular and for meta-analysis in general (Bhoj, 1992; Good, 1955). Although the experimental characteristics or the data collection techniques might suggest possible weighting schemes, e.g. Dempfle and Loesgen (2004), obtaining a set of weights such that the weight for each experiment could be applied uniformly yet meaningfully to thousands of features, as in the present case, is difficult.

Another way to distinguish among experiments is by thresholding them with a pre-specified significance cutoff. Such thresholding has been shown to be effective in cases of criticisms against individual experiments (Darlington and Hayes, 2000). For any feature, only those experimental results which clear the specified cutoff are combined (Olkin and Saner, 2001; Zaykin et al., 2002).

Provided we have knowledge about the quality of the experiments that could lead to a weighting scheme, ESP can be easily generalized as a weighted thresholded product following the routes of Bhoj (1992) and Zaykin et al. (2002). However, common data analysis practices in the microarray bioinformatics community reflect the belief that FDR is good both as a qualitative as well as a quantitative indicator of the overall significance of the experimental results, which are otherwise determined by various factors like the samples, the arrays, the signals and even the statistical tests involved in their generation. Given the heterogeneity that exists among individual investigating teams in terms of their distinct experimental protocols, microarray platforms or scoring schemes, the use of experiment-specific FDR-based cutoffs for meta-analysis by ESP is well justified.

3 THRESHOLDED FISHER PRODUCT

Upon noting that the Fisher product loses power in cases where genuinely significant features have occasional large p-values associated with them, Zaykin et al. (2002) proposed the truncated product method (TPM). In TPM, only those p-values less than or equal to some specified cutoff value \(\tau\) contribute to the product \(W = \prod_{j=1}^{L} p_j^{I(\text{True})}\), where \(I(\text{True}) = 1\) and \(I(\text{False}) = 0\). They also compute the distribution of W and thus the combined probability (see Appendix A).

The cutoff \(\tau\) in the thresholded product W makes it more effective than Fisher’s ordinary product in at least two ways. First, it helps increase the power of the test of the joint null hypothesis by letting some of the highest p-values of a feature be ignored. It is common to find that an otherwise truly significant feature has poor scores in some of the experiments.

Second, a combination of marginally significant p-values might suggest unreasonably high significance (Rosenthal, 1991). TPM guards against such false positives by requiring the presence of at least one p-value significant enough to clear the threshold. In addition, by letting \(\tau = \alpha\), the cutoff also allows such insights as whether the significant entries in the table are indeed significant at the specified level \(\alpha\) (Zaykin et al., 2002).

3.1 Experiment-specific thresholding

Given its usefulness, it seems natural to extend the choice of the cutoff value \(\tau\) such that it is neither arbitrary nor necessarily same for all the experiments, which are not only independent but possibly based on different platforms and statistics. We suggest a more meaningful choice of \(\tau\) guided by the important additional
objective of controlling the FDR in multiple testing of significance in large sets of features from each individual experiment. This addresses the dual concerns of experiment-specific multiple testing and feature-specific meta-analysis in an inter-dependent way.

Genome-wide lists of microarray data are often in need of cutoff values that allow controlling the FDR, which can be computed by several existing algorithms, e.g. Benjamini and Hochberg (1995). Controlling the FDR of the list at \( \alpha \) involves thresholding the list at a feature with \( p \)-value \( \tau_a \) such that if all features in the list up through this one are declared significant, then the FDR would be approximately \( \alpha \).

For multiple experiments \( (2 \leq j \leq L) \), we similarly compute experiment-specific \( p \)-value cutoffs \( \tau_{j,a} \), where the FDR of \( \alpha' \) of an individual experiment is such that the combined FDR satisfies the desired level \( \alpha \).

The \( p \)-value \( p_j \) of a fixed feature for experiment \( j \) is said to ‘clear’ its cutoff \( \tau_{j,a} \) if \( p_j \leq \tau_{j,a} \). Hence the central idea of ESP is to generalize TPM in order to combine only those \( p \)-values of a particular feature which clear their respective experiment-specific FDR thresholding-based cutoffs with the generalized product \( W_a = \prod_{j=1}^L p_j^{p_j/\sum_j p_j} \).

### 3.2 False discovery restrictive thresholds

Reducing Type I error is a primary concern in hypothesis testing. Following Benjamini and Hochberg’s (1995) extension of the idea for multiple testing in terms of FDR, estimation of FDR has become a standard practice in areas like microarray studies. If \( V \) is the number of false positives in the \( R \) features that are declared significant (out of all \( N \) features), then given \( R > 0 \), FDR = \( E(FDP) \) where false discovery proportion, \( FDP = V/R \).

A sequential \( p \)-value method may control FDR as follows: for a chosen \( \alpha' \), it uses some thresholding rule to estimate the index \( k_{\alpha'} \) such that the \( p \)-values \( p_{(1)}, p_{(2)}, \ldots, p_{(k_{\alpha'})} \) may be rejected with FDR level bounded by \( \alpha' \), where \( p_{(1)}, p_{(2)}, \ldots, p_{(N)} \) are the \( p \)-values in ascending order. For instance, the step-up procedure of Benjamini and Hochberg (1995) determines \( k_{\alpha'} \) as the largest \( i \) such that \( p_{(i)} \leq i\alpha/N \). Several stepwise algorithms for FDR control with varying assumptions of dependence among hypotheses are known (Dudoit et al., 2003).

Direct but asymptotic controlling of FDP is also known (Genovese and Wasserman, 2004). Unlike FDR, FDP is controlled in probabilistic sense: as \( \text{Pr}(FDP > \alpha) \leq \gamma \) for chosen \( \alpha, \gamma \) (Genovese and Wasserman, 2002). Recent papers give stepwise algorithms to control FDP which are not asymptotic and work with varied dependence and independence assumptions (Lehmann and Romano, 2005; Romano and Shaikh, 2006; Korn et al., 2004). With simple probability inequalities, a procedure for control of FDR can lead to control of FDP and vice versa (Romano and Shaikh, 2006).

This leads us to our FDR controlling strategy: first, we control the FDP of each individual experiment \( j \) at level \( \alpha \) with the help of cutoffs \( \tau_{j,a} \) derived via corresponding FDR (\( \alpha' \)) control (or possibly directly as \( \tau_{j,a} \), see below) using a known sequential subroutine for the purpose as blackbox. Then we combine across the experiments with respect to their individual FDR controlled thresholds to control the combined FDP at \( \alpha \), and thus the corresponding combined FDR at level \( \alpha \). The relationships among \( \alpha', \alpha, \tilde{\alpha} \) and \( \alpha \) are given in the next section.

### 4 EXTENDING THE THRESHOLDED FISHER PRODUCT

#### 4.1 Ensuring the combined FDR

As noted above, we wish to control the combined FDR at a specified level \( \alpha \) by thresholding each experiment at FDR level \( \alpha' \) (or the corresponding FDP at \( \alpha \)) with experiment-specific \( p \)-value cutoffs \( \tau_{j,a} \) (or directly \( \tau_{j,a} \) as suggested below) for computing the generalized product \( W_a \).

For \( j = 1, 2, \ldots, L \), let \( R_{j,a} \) be the number of features whose \( p \)-values owing to a particular experiment \( j \) clear the corresponding cutoff \( \tau_{j,a} \) and hence declared significant. Then every such feature would have non-trivial product \( W_{a,j} \) owing to at least one experiment in which it is declared significant.

Clearly, if a feature is not found significant in any of the experiments, then it cannot be significant in the combined sense. Hence we pool our combined significant features from the union of significant features from individual experiments. Thus the minimum number of candidate features (\( R' \)) that may appear in the combined significant list \( S_a \) owing to such pooling is given by \( \max_{0 \leq a \leq 1} \sum_{j=1}^L R_{j,a} \).

On the other hand, the actual number of false positives \( V_{j,a} \) for any particular experiment \( j \) is bounded by \( \tilde{\alpha} R_{j,a} \), assuming \( R_{j,a} > 0 \) for all \( j \). Since the false positive features are likely to have low consensus among those declared significant in the independent experiments, there may be as many as \( \sum_{j=1}^L V_{j,a} \) distinct false positives in the combined result. Therefore the maximum proportion of false positives in the combined result is given by \( \sum_{j=1}^L V_{j,a} / R' \leq \sum_{j=1}^L V_{j,a} / R_{j,a} \leq \min (L \times \tilde{\alpha}, 1) \).

This immediately suggests the choice of controlling the FDP of an individual experiment at level \( \tilde{\alpha} = \tilde{\alpha}/L \). Since the experiments are independent and the FDP of every individual experiment \( j \) is controlled at \( \alpha' \), i.e. \( \text{Pr}(FDP > \tilde{\alpha}) \leq \gamma \), the combined FDP is controlled at \( \tilde{\alpha} / L \) owing to \( \text{Pr}(\gamma \cap FDP_j \leq \tilde{\alpha}) > (1-\gamma)^L \) and thus \( \text{Pr}(FDP > \tilde{\alpha}) \leq 1 - (1-\gamma)^L \leq \gamma \) for the choice of \( \gamma = \gamma/L \).

By Markov’s inequality, if for every experiment \( j \), FDR \( FDP_j \leq \alpha' = (\alpha/L)(\gamma/L) \), then \( \text{Pr}(FDP_j > \tilde{\alpha} / L) \leq \gamma/L \). By the above argument, this implies \( \text{Pr}(FDP > \tilde{\alpha}) \leq \gamma \) for the combined result. As shown in Lehmann and Romano (2005), this leads to combined FDR \( \leq \gamma(1 -\alpha) + \alpha \leq \alpha \) for the choice of \( \alpha = \alpha + \gamma \).

If \( L \) and \( \alpha \) are fixed, then the largest value of \( \alpha' \), owing to \( \tilde{\alpha} = \gamma / 2 \), is \( \alpha' = 2(1/L)^2 \). Clearly, even this large value admits of a very conservative FDR control owing to the crudeness of Markov’s inequality, and one can alternatively begin with control of experiment-specific FDP with cutoffs \( \tau_{j,a} \) instead of FDR.

For instance, for \( L = 3 \) and \( \alpha = 0.06 \), possible FDP parameters are \( \tilde{\alpha} = 0.01 \) and \( \gamma = 0.01 \), although one might want to choose the values such that \( \tilde{\alpha} \approx 0.1 \).

Hence, for any fixed feature and a chosen combined FDR level \( \tilde{\alpha} \), by letting \( \tilde{\alpha} = \tilde{\alpha}/L \), the combined probability distribution for the product \( W_a = \prod_{j=1}^L p_j^{p_j/\sum_j p_j} \) can be generalized from that of TPM in a straightforward manner (Appendix A).

#### 4.2 The consensus parameter

We note that the above thresholding works in exactly similar way if we use alternatively experiment-specific cutoffs \( \tau_{j,a,\gamma} \) which control the number of false positives (\( Ip \)) instead of proportion FDP (Korn et al., 2004). Clearly, allowing at most \( \tilde{\alpha} \) low-consensus false positives due to each of the \( L \) experiments should yield a tight upper bound of \( \tilde{\alpha} = \tilde{\alpha}/L \) on the total number of combined false positives (with probability \( \gamma = \gamma/L \) earlier).
Further, the control of $fp$ allows us to use a global consensus parameter $C$ (an integer between 1 and $L$) which requires that to be declared significant in the combined sense, a feature must be significant in at least $C$ experiments. The consensus constraint can be similarly applied to the FDP thresholded pool of combined significant features described in the last section provided the size of the constrained pool $S_{a,c}$ is, as earlier, at least $\max_j \{1, 2, \ldots, L\} R_j a^c$.

Thresholding the $p$-values of a feature and ensuring the number of $p$-values of a feature that need to clear their cutoffs in order to allow meaningful meta-analysis are dual concerns. The consensus parameter $C$ thus imposes quality control on the combined significant set, which becomes obvious even with $C=2$ that guards against noise spikes which are unlikely to repeat for the same feature. Yet, the choice of $C$ should follow from a moderate vote-counting rule, like $\frac{1}{2}L$ (Hedges and Olkin, 1985), so as not to unnecessarily filter true positives and thus affect both FDP and power.

If $L$ is large enough to merit $O(M)$ pre-processing (say, $L > \log_2 N$), then for the sake of speeding up the computation of combined probabilities, for every feature $g$, we can opt to pre-compute (i.e. before specification of $a$) its bounds $\mathcal{C}_g$ and $\mathcal{U}_g$ (integers between 1 and $L$) such that the probability of fewer (or more) than $\mathcal{C}_g$ (respectively $\mathcal{U}_g$) $p$-values clearing their respective randomly chosen cutoffs is less than a specified global constant $p_{\text{search}} \in [0, 1]$. A slow exact method followed by a fast approximate alternative (using probabilistic tail bounds) to do this computation are suggested in Appendix B. These parameters reduce an otherwise necessarily exponential computation time [Equation (4), Appendix A] by approximating the combined probability with a conditional sum [Equation (5)] that is defined only over sets of $K$ $p$-values which might clear their corresponding cutoffs, i.e. for $\mathcal{C}_g \leq K \leq \mathcal{U}_g$.

5 EXPERIMENTAL RESULTS

5.1 Simulation studies

As a first evaluation of ESP, we simulated a population with $N = 10^4$ $p$-values. A proportion $\nu$ of the population was comprised of features with significant $p$-values. These significant $p$-values are chosen from the tail of a distribution such that 99% of the $p$-values were <0.05 (see Appendix C for details), our chosen significance level. The remainder of the population was comprised of features with $p$-values chosen randomly from a uniform distribution between 0 and 1. From this ‘original’ population $\mathcal{O}$, we then generate $L = 5$ experiments by adding Gaussian noise to the $p$-value of each of the features (Appendix C), and these $L$ experiments were subjected to meta-analysis by ESP, the standard Fisher product and other similar methods.

At a given FDR level ($\alpha = 0.05$), the features found to be significant by ESP ($S_M$), and the features found to be significant by the Fisher method ($S_F$) were then compared with the features of the original population that were significant ($S_O$) at the same level (Fig. 1). The above process of generating $L$ experiments from the population $\mathcal{O}$ and then conducting meta-analysis was repeated for $10^4$ simulation runs, and results were averaged (Fig. 1). Finally, the whole process was repeated on 20 different original populations, with $\nu$ ranging from 0.005 to 0.1. The results are shown in Figure 1.

First [Fig. 1a], we report the ratio of the number of significant features found by ESP to the number of significant features in the original population (i.e. $|S_M|/|S_O|$), and we report this ($y$-axis) not only as a function of $\nu$, but also as a function of $C$, the consensus parameter. A significant deviation from the line $y = 1$ either indicates more false positives or less power (although the converse is not true). Therefore, the choice of $C=2$, whereby the frequency of $S_M$ almost equals that of $S_O$ as $\nu$ increases [Fig. 1a], is likely to allow good FDR control without loss of power. The issue of significance of the features in $S_M$ obtained by ESP with $C=2$ is addressed below.

We note that many features are likely to have no significance in all simulation runs, and results were averaged (Fig. 1a). Finally, the choice of $C=2$ should follow from a moderate vote-counting rule, like $\frac{1}{2}L$ (Hedges and Olkin, 1985), so as not to unnecessarily filter true positives and thus affect both FDP and power.

![Fig. 1. Simulation results obtained over $20 \times 10^3$ runs. The x-axis in all the plots represents the proportion $\nu$ (in the range [0.005, 0.1]) of significant features in the original distribution for each run. The suffix letters ‘F’, ‘R’, ‘S’ and ‘M’ denote Fisher product, RTP, Sidak and ESP respectively. The $y$-axes for plots (a)–(c) represent (a) the ratio of significant features found by ESP to significant features in the original population, i.e., $(|S_M|/|S_O|)$, calculated for different values of $C = 1 \ldots L$, (b) rank correlation of significant features $S_M$, $S_F$, $S_R$, $S_S$ and $S_O$ with the corresponding $p$-values in $O$ and (c) the overlap of the significant features with $S_O$. Throughout the figure, $\alpha = 0.05$, and in plots (b) and (c), $C=2$.](https://academic.oup.com/bioinformatics/article-abstract/22/20/2516/218866)

by guest

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/22/20/2516/218866 on 10 April 2018
is well below the allowed $\alpha$ and demonstrates strong and conservative FDR control. Indeed its closeness to the experiment-specific threshold (provided $\alpha = \bar{\alpha}/L \approx \alpha/L$) indicates the effectiveness of the consensus criterion in cross-experimental filtering of the experiment-specific noise.

We also report the following measures—(1) the proportions $\phi$ of significant features by the various combination methods (e.g. $S_{SM}$ and $S_{SF}$) which are also significant in the original distribution $O$ at the same level $\alpha$ and (2) the rank correlations $\rho$ of the features in $S_{SM}$ and $S_{SF}$ with the corresponding $p$-values in $O$—thus $\rho$ and $\phi$ together act as an enrichment measure for the output of meta-analysis. The results are shown in Figure 1b and c.

When $\nu$ is small, ESP performs significantly better than Fisher: the few significant features present in $O$ are lost to noise and go undetected by Fisher product, but the same are identified by the false discovery restrictive thresholds of ESP. For all values of $\nu$, a total of 80–90% of the features in $S_{SM}$ are found to be significant in $O$ along with an overall correlation of almost 0.9 (Fig. 1b and c). In contrast, although Fisher product shows comparable correlation for higher values of $\nu$ (Fig. 1b), in the absence of thresholding and consensus, the percentage of features in $S_{SF}$ that are actually significant in $O$ (i.e. overlapping with $S_{SM}$) never exceeds 50% (Fig. 1c).

We also used other comparable techniques like rank truncated product [or RTP, Dudbridge and Koeleman (2003)] of $K$ smallest of the $L$ $p$-values, and its special case with $K = 1$ that corresponds to Sidak’s correction (Sidak, 1967), to combine the simulated data. To validate the effectiveness of our consensus parameter $C = 2$, we chose to do RTP with $K = 2$. Clearly, ESP yields more information-rich products for the significant features than RTP (hence also Sidak’s) since these products may be formed not only of more than two $p$-values but which have also cleared their cutoffs. The effectiveness of ESP compared with other techniques is demonstrated in the higher correlation and significance measurements in Figure 1b and c.

We performed a further simulation study with microarray-type data distributions to explore the effects of varying noise. Here we used two kinds of noise: we added small amounts of Gaussian noise to each member of the population, but in addition, we added large bursts of noise to a small randomly chosen subset of the population. We believe that this mimics the situation in actual microarray experiments, e.g. a ChIP-chip study. Again, ESP out-performed the Fisher product by a wide margin. See Appendix D for details.

5.2 ChIP-chip studies

Proteins called transcription factors (TFs) regulate transcription by binding to DNA motifs upstream of their target genes. The availability of the genome sequence of budding yeast (Saccharomyces cerevisiae) allowed chromatin immunoprecipitation (ChIP) (used to identify protein–DNA interactions) to be coupled to high-throughput analysis on microarrays, or ‘chip’-s, to monitor and measure the binding of a given set of TFs to the upstream regulatory regions of thousands of genes—this is referred to as a ‘ChIP-chip’ experiment.

The present algorithm, implemented with custom code written in Perl, was applied to combine three well-known ChIP-chip genome-wide TF binding datasets to get a consensus result. These datasets were all generated in the laboratory of Dr R. Young at the Whitehead Institute using similar procedures and reagents. Because of this, these three datasets are already reasonably similar to each other, in comparison with datasets that might have been generated in three distinct labs. This pre-existing homogeneity limits the improvement that our algorithm (or any other similar algorithm) can achieve. However, these datasets provide a good test case because there are other comparable or relevant datasets (Iyer et al., 2001; Spellman et al., 1998) which can be used to address the final correctness of the original or combined results.

We index the upstream intergenic regions in the datasets by the possible transcriptional target ‘genes’ to obtain a combined list of 6401 genes over three experiments given by the sets denoted $L$ (Lee et al., 2002), $S$ (Simon et al., 2001) and $H$ (Harbison et al., 2004) containing $p$-values which measure the binding of different TF proteins to the upstream regions of the said genes. The missing entries are marked in ESP with a $p$-value of 1. Meta-analysis results for the protein Swi4 that forms a part of the transcription factor SBF are presented below, since these could be validated with the help of shortlisted 208 genes for Swi4 (and MBF) owing to a fourth comparable dataset I (Iyer et al., 2001). With $C = 2$, significant intergenic regions for 106 genes have been identified for Swi4; this list and similar results for eight other TFs, all of which are known to regulate the yeast cell cycle as studied in Simon et al. (2001), are listed in the Supplementary Materials website.

For reporting our results, the combined FDR level $\alpha$ is chosen to be 0.06. The $q$-value for a particular feature is defined in Storey and Tibshirani (2003) as the minimum expected proportion of false positives occurring up through that feature on the list. The divergent $q$-values of the most significant 210 genes in the three datasets are plotted in Figure 2, which justifies the use of distinct $p$-value cutoffs by our algorithm. When the same cutoffs for $\alpha = 0.06$ are applied

![Fig. 2. The estimated false discovery rates for 210 most significant genes from each of the original datasets H, S and L. The divergence in the three plots highlights the need for different experiment-specific cutoffs for thresholded Fisher product.](https://academic.oup.com/bioinformatics/article-abstract/22/20/2516/218866)
to the individual datasets L, S and H, they yield significant subsets of 110,137 and 131 genes respectively.

Since L is only 3, we do the meta-analysis with the global parameters U = 3 and C separately for all values 1, 2 and 3. Although the global parameter C allows regulation of the degree of consensus between 1 (set union), 2 (2-to-1 majority rule) and 3 (set intersection), we choose C = 2 to report our output. Thus our algorithm selects a set of 106 genes as significant for the TF Swi4. Numbers of significant genes similarly selected for different levels of FDR and for all values of C are given in Table 1.

To assay the success of our meta-analysis, we compute the Spearman rank correlation of our output list of significant genes (ranked by their combined probabilities) with the list I, and then compare this with four rank correlations of I with (p-value) ranked lists owing to four control conditions. In all cases, the rank correlation of two lists is computed with only the genes that are common to both. Of the above 106 genes in our output list, 67 are shared with I and the rank correlation is 0.71.

In contrast, ordinary Fisher product combination of L, S and H in terms of the top 67 genes gives a lower correlation of 0.65. Without the consensus requirement, i.e. for C = 1, although the number of significant genes shared with I increases from 67 to 80 (out of a total of 171; Table 1), the correlation drops from 0.71 to 0.64.

Without the consensus available to our algorithm, at the same FDR level, such correlations with I for each of the standalone datasets may expected to be lower than that for the combined data. However, the similarity inherent to the three experiments limits the scope of distinguishing between the original and the combined results, particularly at the selection range of the most significant genes. Thus the rank correlations of I and the individual datasets H, S and L are, not surprisingly, 0.67, 0.72 and 0.70, respectively. These are computed with the 67, 65 and 66 genes that are common to I and H, S and L, respectively.

According to the percentile ranks given by I, the 67 genes that I shares with our output set have their median percentile rank as high as 0.9904, which corresponds to the top ranked 20% genes in I; the gene set enrichment validates the significance of the genes in the output of ESP. Finally, as a more radical control condition, if we use the ranked list for the transcription factor MBF (MBF is a heterodimer of TFs Swi6 and Mbp1) also available from I, then the correlation with our output set (i.e. for Swi4, which does not form a part of MBF) is found to be only 0.02.

### Table 1. Number of significant genes for swi4 at different values of FDR and C

<table>
<thead>
<tr>
<th>FDR α</th>
<th>All 3</th>
<th>At least 2</th>
<th>At least 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>15</td>
<td>36</td>
<td>74</td>
</tr>
<tr>
<td>0.005</td>
<td>42</td>
<td>58</td>
<td>95</td>
</tr>
<tr>
<td>0.01</td>
<td>48</td>
<td>65</td>
<td>108</td>
</tr>
<tr>
<td>0.025</td>
<td>59</td>
<td>84</td>
<td>139</td>
</tr>
<tr>
<td>0.05</td>
<td>64</td>
<td>103</td>
<td>162</td>
</tr>
<tr>
<td>0.06</td>
<td>65</td>
<td>106</td>
<td>174</td>
</tr>
<tr>
<td>0.075</td>
<td>70</td>
<td>113</td>
<td>183</td>
</tr>
<tr>
<td>0.1</td>
<td>76</td>
<td>125</td>
<td>199</td>
</tr>
</tbody>
</table>

See Supplementary Materials for other TFs.

### 6 Discussion and Conclusion

We present here a novel algorithm to integrate the critical issues of meta-analysis and the control of FDR. In the process we introduced several concepts, namely (1) use different cutoffs in the thresholded Fisher product, (2) base these cutoffs on experiment-specific false discovery restrictive thresholds, (3) optionally impose quality control on the meta-analysis with a consensus criterion and (4) optionally speed up the algorithm by limiting the search with the help of probabilistic tail bounds to those subsets of p-values that are most likely to yield the product.

Applications for the present algorithm need not be restricted to microarray datasets. While considering synthesis of large amounts of genome-wide data, particularly if derived from heterogeneous technologies and statistics as they often are, the use of different cutoffs should prove to be natural and insightful whenever it is required to combine significance values. For instance, this may be applied to the product of match probabilities used for multiple motif analysis (Bailey and Gribskov, 1998). The false discovery restrictive thresholding offered by ESP is an added built-in advantage that guards against defective entries and false positives.

We note that the consensus parameter offers its own additional (cross-experimental) control against false discoveries and this may result in conservative combined control as seen in the simulation of Section 5.1. Also the experiment-specific control at α = α / L, particularly when L is large, could prove to be stringent. Thus, we want to further extend the meta-analysis procedure by relaxing the experiment-specific control to be at level α / K for the largest K in C ≤ K ≤ L such that the product for a feature may be formed of K of its smallest p-values which clear their cutoffs (this lets the feature be represented in terms of its most significant entries) while the combined FDR is observed. This would generalize the ranked truncated extension of TPM owing to Dudbridge and Koeleman (2003), although being not restricted to a single common cutoff (τ), the combined probability computation in this case may not make direct use of Beta distribution as in their article.

The present article describes a method for synthesis and subsequent extraction of the most significant subsets of features from large and noisy datasets for further investigation. This could prove helpful for a large number of bioinformatics applications such as synthesis of cell cycle experiments, cancer microarray data and various high-throughput screens (Oliva et al., 2005; Rhodes et al., 2004; Grutzmann et al., 2005; Choi et al., 2003). Besides identification of significant subsets of features like gene modules related to cancer, it can naturally lead to recognition of important patterns and pathways discovered only by considering different studies together.

### Acknowledgements

The authors thank the reviewers for their helpful suggestions. This work was supported by NIH grant GM064813 and NSF grants EIA-0325123 and DBI-0444815.

### Conflict of Interest

none declared.

### References


