BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 2 2012, pages 206–213
doi:10.1093/bioinformatics/btr628

*Gene expression*

Advance Access publication November 13, 2011

# Discovering transcription factor regulatory targets using gene expression and binding data

Mark Maienschein-Cline[1], Jie Zhou[2], Kevin P. White[2,3], Roger Sciammas[4] and Aaron R. Dinner[1,3,*]

[1]Department of Chemistry, [2]Department of Human Genetics, [3]Institute for Genomics and Systems Biology and [4]Department of Surgery, The University of Chicago, Chicago, IL 60637, USA

Associate Editor: Janet Kelso

**ABSTRACT**

**Motivation:** Identifying the target genes regulated by transcription factors (TFs) is the most basic step in understanding gene regulation. Recent advances in high-throughput sequencing technology, together with chromatin immunoprecipitation (ChIP), enable mapping TF binding sites genome wide, but it is not possible to infer function from binding alone. This is especially true in mammalian systems, where regulation often occurs through long-range enhancers in gene-rich neighborhoods, rather than proximal promoters, preventing straightforward assignment of a binding site to a target gene.

**Results:** We present EMBER (Expectation Maximization of Binding and Expression pRofiles), a method that integrates high-throughput binding data (e.g. ChIP-chip or ChIP-seq) with gene expression data (e.g. DNA microarray) via an unsupervised machine learning algorithm for inferring the gene targets of sets of TF binding sites. Genes selected are those that match overrepresented expression patterns, which can be used to provide information about multiple TF regulatory modes. We apply the method to genome-wide human breast cancer data and demonstrate that EMBER confirms a role for the TFs estrogen receptor alpha, retinoic acid receptors alpha and gamma in breast cancer development, whereas the conventional approach of assigning regulatory targets based on proximity does not. Additionally, we compare several predicted target genes from EMBER to interactions inferred previously, examine combinatorial effects of TFs on gene regulation and illustrate the ability of EMBER to discover multiple modes of regulation.

**Availability:** All code used for this work is available at http://dinner-group.uchicago.edu/downloads.html

**Contact:** dinner@uchicago.edu

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

Received on August 22, 2011; revised on November 4, 2011; accepted on November 8, 2011

## 1 INTRODUCTION

The fundamental step in gene regulation is a transcription factor (TF) binding DNA. Recent advances in DNA sequencing and microarray technologies have led to huge increases in the amount of information that biologists can accumulate about this process (Berger and Bulyk, 2006; Capaldi *et al.*, 2008; Johnson *et al.*, 2007; Kim *et al.*, 2009; Pepke *et al.*, 2009; Robertson *et al.*, 2007). In particular, ChIP-chip and ChIP-seq experiments have enabled genome-wide mapping of TF binding sites in a single experiment. Knowledge of these binding sites is one of the key steps in determining which genes are regulated by a TF, but the binding sites alone are not sufficient to infer regulation.

In simple organisms, such as bacteria or yeast, most transcriptional regulation occurs by a TF binding in a promoter region, near the transcription start site (TSS) of a gene (Capaldi *et al.*, 2008; Ptashne, 1992). However, as organisms increase in complexity, more regulation occurs through long-range enhancers, often spanning many tens of thousands of base pairs (Arnosti and Kulkarni, 2005). In fact, these enhancer interactions become the principal form of gene regulation in more complex species. For example, in the human ChIP-chip data (Hua *et al.*, 2008, 2009) that we discuss later in this article, only 5–10% of the binding sites are in promoters (defined as 2 kb from a TSS). In this case, it is not obvious which gene a particular binding site may be regulating: enhancer regulation is independent of the binding site orientation relative to the regulated gene (Banerji *et al.*, 1981), and even transcription factors bound to a particular gene may regulate a different gene through an enhancer (Lee *et al.*, 2005). Additionally, binding sites in gene-rich locations of the genome have a large number of potential targets nearby. Although we may have some idea of the maximum range at which an enhancer is likely to function (Arnosti and Kulkarni, 2005), any gene within that distance must be considered a potential target.

A common approach to assigning gene targets to TF binding sites is simply to use proximity: regulation is inferred by the presence of a binding site close to a gene, most often by assignment to the nearest gene (Fujiwara *et al.*, 2009; Gilchrist *et al.*, 2006; Verzi *et al.*, 2010; Yu *et al.*, 2009). This approach is clearly an over simplification, as it is not clear that a gene 30 kb from a gene is significantly more likely to be regulated than a gene 40 kb away, for example. More importantly, this approach can miss important long-range interactions and is unable to distinguish functional from non-functional binding sites. In situations where enhancer regulation is predominant, more information is needed to properly pair binding sites with their regulatory targets.

We propose using gene expression information to assign TF binding sites to gene targets. Namely, we assume that genes regulated by a TF should behave in some measurably consistent manner. That is, there should be some subset of true targets within the many potential targets surrounding each binding site, and we should

*To whom correspondence should be addressed.

be able to distinguish these true targets by their overrepresented behaviors. Although there may actually be some variations in regulatory behavior within the TF binding sites, such as when TFs act combinatorially, this general idea is a useful motivation for a data-driven target identification method.

There have been several studies published that explore the integration of gene behaviors with TF binding data. One recent study (McLean *et al.*, 2010) uses gene ontology, assigning gene targets by looking for ontological classifications that are overrepresented among a subset of the potential target genes. Although this approach does not require any additional experiments, it is limited by the availability and accuracy of gene ontology databases and lacks the quantitative aspect that additional experimental data would provide. A number of other methods use DNA microarray experiments to quantify gene expression. Many of these methods focus on the identification of TF activity profiles, aiming to identify subsets of binding sites where the regulated genes behave in a highly correlated fashion, and thus are co-regulated (Bar-Joseph *et al.*, 2003; Boulesteix and Strimmer, 2005; Gao *et al.*, 2004). Others use a modeling approach, fitting dynamical networks to the observed behaviors in order to infer gene regulation (Sanguinetti *et al.*, 2003, 2006). However, these methods focus on promoter regulation, primarily whether binding sites are functional or not. In contrast, our goal is to identify which gene, out of a number of possibilities, is regulated by an enhancer site.

In this article, we propose a particular integration of TF binding and gene expression data paired with an expectation maximization algorithm (Bailey and Elkan, 1994) for unsupervised machine learning, to simultaneously discover the gene regulatory targets of a set of TF binding sites and the expression pattern exhibited by the regulated genes. We call the method expectation maximization of binding and expression pRofiles (EMBER). Since EMBER quantitatively ranks all potential target genes based on a discovered expression pattern, it can infer whether binding sites are functional and, in turn, their regulatory targets. Also, EMBER can search for more than one expression pattern, allowing for the discovery of multiple regulatory modes. The combinatorial effect of multiple TFs can be studied as well by searching for expression patterns among different overlapping peak sets.

In Section 2, we first describe our data integration approach. Then we motivate and describe the EMBER algorithm. In Section 3, we apply the method to ChIP-chip and DNA microarray data (Hua *et al.*, 2008, 2009) probing the role of TFs estrogen receptor alpha (ER$\alpha$), retinoic acid receptor alpha (RAR$\alpha$) and retinoic acid receptor gamma (RAR$\gamma$) in human breast cancer development. In our primary result, we find a strong bias toward breast cancer-related genes for the targets of all three of the above TFs, consistent with their involvement in breast cancer development, and we observe that the conventional proximity approach does not obtain this result. Additionally, we compare the results of EMBER to inferences made in a previous publication of the data (Hua *et al.*, 2009). Specifically, we confirm the regulation of *FOXA1*, *FOS* and *GATA3* by these TFs, as well as regulation of *ESR1* by the RAR factors, although we do not observe regulation of either *RARA* or *RARG* by these TFs, as had been suggested. We also investigate the combinatorial effects of these TFs, in particular, noting some different regulatory behaviors by ER$\alpha$ in combination with RAR$\gamma$. Finally, we illustrate EMBER's ability to find multiple regulatory modes for a single set of binding sites, and demonstrate its robustness with respect to the range over which genes are considered.

## 2 METHODS

Before describing the target identification algorithm in EMBER, we lay out the context of the problem and give a general overview of the steps involved. We start with a series of experiments measuring expression and at least one experiment measuring binding. DNA microarray data should be normalized [by any of the standard procedures used for microarrays (Li and Wong, 2003; Smyth, 2004; Townsend and Hartl, 2002; Wu *et al.*, 2003)] and curated to remove control probe sets and ones that are never expressed to a high level. ChIP-seq or ChIP-chip data should be aligned with peaks called [again, by standard procedures (Bernstein *et al.*, 2005; Cawley *et al.*, 2004; Pepke *et al.*, 2009; Valouev *et al.*, 2008)]. Given these data, there are essentially four steps, which are illustrated in Figure 1. The first three steps are preprocessing and data integration steps, mostly in regards to the DNA microarray data. These initial analyses are detailed in the next subsection. In brief, the steps are (i) to define behavior dimensions from binary comparisons between the microarray conditions (Fig. 1A); (ii) to classify each gene's behavior in each dimension and compile the resulting behavior profiles for the entire microarray (Fig. 1B); and (iii) to integrate the expression data with the TF binding data by choosing potential regulatory targets for each binding site (Fig. 1C). The final step is to run EMBER and assign targets based on its results (Fig. 1D). The details of the algorithm are given later in this section and in Section 1 of the Supplementary Materials, and the outputs of the method are described and illustrated as well.

### 2.1 Preparation and integration of data

We focus first on the gene expression data, Part I of Figure 1, measured by DNA microarray. These experiments are typically performed in series over a variety of cellular conditions, such as a time course over treatment with some chemical, or in wild-type and mutant cells. In our approach, we summarize the change in each probe set's expression over the series of microarray conditions into a discrete space, taking into account the natural range in expression levels for each probe set at each condition. In principle, the expression patterns could be constructed from the behaviors of genes, rather than probe sets. We choose to work directly with the probe sets to avoid imposing an arbitrary combination rule when the probe sets for a gene are divergent. For the breast cancer data that we analyze in the present article, only one probe set per gene remained after removing low-expressed probes, so probe sets and genes are equivalent in this case, and we use the terms interchangeably below.

First, we define a series of *behavior dimensions* by considering appropriate binary comparisons between different DNA microarray conditions. For example, in Figure 1A if a series of experiments covered a time course after treatment with some chemical (time 0, time 1, time 2, time 3), then three behavior dimensions might be time 0 to 1, time 1 to 2 and time 2 to 3. The choice of the experiments is important because including expression data from cell states very different from the binding data can obscure the relevant information. Relative entropy analysis of the EMBER results (Section 3.4) gives some insight into the expression comparisons that are most relevant, but it is ultimately the expertise of the researcher that determines the appropriate match of binding and expression data.

Second, we discretize each probe set's behavior in each of the above dimensions, illustrated in Figure 1B. Since we assume that we have several replicates for each condition, we can calculate a mean ($\mu$) and standard deviation ($\sigma$) of expression for each probe set at each condition; we denote the means/standard deviations for the first and second conditions of a comparison as $\mu_1/\sigma_1$ and $\mu_2/\sigma_2$, respectively. Behaviors are then classified in each dimension into one of five values, $++, +, 0, -$ or $--$, denoting large upregulation, small upregulation, no change, small downregulation or large downregulation, respectively. These are defined by considering the differences between the means, relative to the magnitudes of the sum of

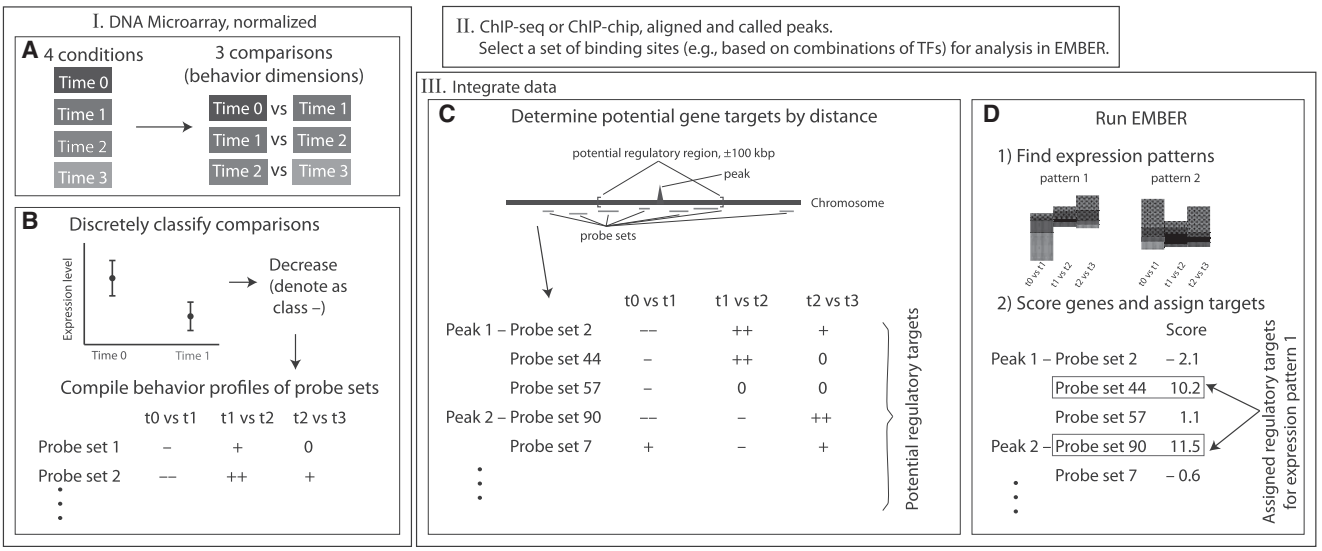


**Fig. 1.** Schematic of the data preparation and integration steps. Part I: DNA microarray preparation, starting with data normalized by standard procedures, for a hypothetical series of four microarray experiments. (**A**) Different conditions lead to the definition of different binary behavior dimensions. (**B**) The behavior of each probe set in each dimension is classified by considering the difference in mean expression levels, relative to the standard deviations (cartoon depicts the mean as a dot, with error bars signifying $\pm 1$ SD), as described in the text. Binding sites obtained from standard ChIP-seq or ChIP-chip alignment and peak calling (Part II) are integrated with the gene expression data (Part III). (**C**) Peaks are assigned potential targets by considering all genes within 100 kb of the binding site. In the figure, this region is given by the brackets around the peak. The orange probe sets, which lie within this region, are assigned as potential targets, and the green probe sets are not. This results in a new data structure of integrated binding and gene expression data. (**D**) These data, together with the background from (B), are used to (1) identify overrepresented patterns in EMBER, and the resulting expression patterns are used to (2) score probe sets and assign regulatory targets.

the standard deviations, $s = \sigma_1 + \sigma_2$. We define a large upregulation (++) as $3s < \mu_2 - \mu_1$ (which implies $\mu_2 > \mu_1$ as well, indicating an increase in expression). A small upregulation (+) is $s < \mu_2 - \mu_1 \leq 3s$. No change (0) is $-s \leq \mu_2 - \mu_1 \leq s$. Downregulation classifications (−− and −) are defined analogously to upregulation classifications by reversing the order of $\mu_2$ and $\mu_1$. This discretized behavior profile facilitates detection of overrepresented patterns. We also prefer this approach to the more common analysis of fold change in that it incorporates our confidence in the change in expression. Note that our behavior profile definition is similar in spirit to other approaches to analyzing DNA microarray data, as this transformation allows for clustering (Eisen *et al.*, 1998; Gilchrist *et al.*, 2006; MacQueen, 1967) or the definition of modules (Bar-Joseph *et al.*, 2003) in behavior.

We now integrate the TF binding data. We define a 'potential regulatory region' (Fig. 1C) for each binding site of $\pm 100$ kb. Any probe set whose gene overlaps with this region is assigned as a potential target. This distance was chosen as a compromise between including as many potential targets as possible and maximizing signal to noise. We have found the results of EMBER to be robust to variations in this parameter; see Section 2 and Figure 1 of the Supplementary Materials for details. The result is a number of potential regulatory targets for each binding site. Our problem is thus to identify which of these are actually being regulated by binding of a particular set of TFs.

## 2.2 The EMBER method

Our problem of identifying regulatory targets can be mapped onto the common problem of identifying precise TF binding sites within binding regions obtained from ChIP-seq or ChIP-chip experiments. In the latter problem, the experimental method yields TF binding sites to a precision of a couple of hundred base pairs, but the actual site is expected to be only 10–20 bp. To identify the actual binding sites, one starts with a large number of potential binding sites (e.g. all possible DNA 10mers from the ChIP regions) and searches among those sites for a set of overrepresented DNA sequences that identifies the putative binding sites. The resulting collection of sequences is called the binding *motif*.

Analogously, in identifying regulatory targets we assume that a given TF regulates genes in some consistent fashion, reflected in the gene expression profiles. Starting with a large number of potential targets (i.e. all genes in the potential regulatory region, $\pm 100$ kb from the peak), we search among those genes for overrepresented behavior profiles. We refer to the resulting collection of gene behavior profiles as the *expression pattern*.

There are, however, a few important differences between binding motifs and expression patterns. The backgrounds for the two problems are quite different: in the TF binding site identification problem, the background comes from the background frequency of each DNA base and is invariant over the length of the motif. The background frequencies for A, G, T and C fall in the range 17–33%, depending on the organism. For the gene targeting problem, the background is much farther from uniform (generally the largest frequency is at the '0' classification, as most genes do not have appreciably different expression levels under comparison of two conditions). Additionally, in an expression pattern, each comparison has a different background, based on all the genes in the microarray.

Gene behavior profiles also cannot overlap, in contrast to TF binding sites. An important consequence of this difference is that the space of possible expression patterns is more constrained than the space of potential TF binding sites; we have found that the converged patterns are completely independent of the initial conditions used, which is not generally true of binding motifs.

There are a number of algorithms available to identify overrepresented patterns, and in principle many of them could be adapted for our target identification problem. Here, we adapt the MM expectation maximization algorithm of MEME (Bailey and Elkan, 1994), detailed in Section 1 of the Supplementary Materials.

In brief, the goals of the method are (i) to separate the potential targets into target and background genes, and (ii) to determine multinomial probability

models for the target and background genes such that the likelihood of observing the gene behaviors under the models is maximized. The essential parameters that are found in an expression pattern are the probabilities in the models, $f_{j,l,m}$ and $\lambda_j$. Here, $j=1$ for the target genes and $j=2$ for the background genes; $l$ indexes the behavior dimensions (e.g. time 0 to time 1), and $m$ indexes the behavior classifications (e.g. $++$). $f_{j,l,m}$ is the probability of observing behavior $m$ in comparison $l$ for the $j$-th model, and $\lambda_j$ is the *a priori* probability of a potential target gene belonging to model $j$.

After the convergence of each expression pattern, targets are assigned based on a score $S(\mathbf{x}_i)$:

$$S(\mathbf{x}_i)=\sum_{l=1}^{L}\sum_{m=--}^{++}\log(f_{1,l,m}/f_{2,l,m})\delta_{m,x_{i,l}}. \quad (1)$$

The score of gene $\mathbf{x}_i$ quantifies how much more the gene belongs to the expression pattern model than the background model. A natural threshold for assignment is $T=\log(\lambda_2/\lambda_1)$.

In addition to determining the target genes, it is can be useful to directly examine the log-odds scoring matrix with elements $S_{l,m}=[\log(f_{1,l,m}/f_{2,l,m})]$, hereafter referred to as the score matrix. Positive values in the score matrix represent behaviors that are overrepresented in the expression pattern, and negative values represent behaviors that are underrepresented. One can also compare score matrices to quantify how two expression patterns differ. This is particularly useful when seeking to elucidate the actions of different combinations of TFs; even if the sites, and hence the target genes, differ significantly, the score matrices may be quite similar, implying a similar mode of regulation.

## 3 RESULTS

### 3.1 Data used for analysis

We demonstrate EMBER by applying it to the previously published (Hua *et al.*, 2008, 2009) data for the TFs ER$\alpha$ and RAR$\alpha$ and RAR$\gamma$ in breast cancer MCF-7 cells. ChIP-chip binding data were collected for these three TFs, and DNA microarray data were collected at different time points after treatments with estrogen (E2) and different RA agonists (ATRA, which binds to both RARs; AM580, RAR$\alpha$ specific; CD437, RAR$\gamma$ specific; and a combination of AM580 and CD437, called AM + CD). A summary of the data (including the behavior dimensions defined) is given in Supplementary Figure 2A and B, respectively.

We prepared the data as described in Section 2. We separately applied EMBER for all the binding sites for each TF, as well as the binding sites in each of the seven coincident binding regions in the Venn diagram in Supplementary Figure 2A, for a total of 10 EMBER runs (each run takes $\sim$1 min; binding sites are called coincident if the center of two ChIP-chip binding regions are within 1 kb). In other applications, it could be of interest to compare the cistrome of the same TF(s) at different time points to study how its regulation changed with cell state. Our analysis in this section focuses primarily on the gene targets obtained in the top expression pattern in each run.

Using these data, we can also more clearly illustrate the target-gene finding problem. For example, only 7.7% of all RAR$\alpha$ peaks are within a promoter region (defined as 2 kb from a TSS), so most sites regulate genes through long-range enhancers. Supplementary Figure 3 looks at the distribution of potential targets (within 100 kb) per peak: a substantial number of peaks have at least a handful of nearby genes, and some have as many as 29. Although it is a common approach in these cases to assign targets by proximity, we will show that this method can miss important information that is obtained in EMBER.

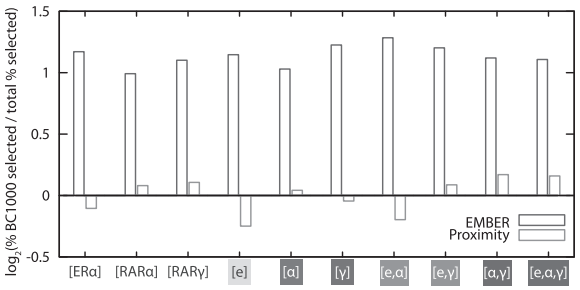


**Fig. 2.** Bias toward BC1000 genes in EMBER versus proximity. Plotted is the $\log_2$ ratio of the fraction of potential target BC1000 genes that are selected as targets, to the total fraction of genes that are selected as targets, for each method.

### 3.2 Overrepresentation of breast cancer-related genes in EMBER-identified targets

We first investigated whether we could link the target genes obtained from EMBER to the biology of ER$\alpha$, RAR$\alpha$ and RAR$\gamma$. Estrogen (E2) is involved in initiation, development and metastasis (Yager and Davidson, 2006) of breast cancer, and RA can hinder the proliferation of tumor cells (Liu *et al.*, 1998; Zhou *et al.*, 1997). Both E2 and RA exert genomic effects through their activation of ER$\alpha$, and RAR$\alpha$ and RAR$\gamma$, respectively. Hua *et al.* (2009) previously manually curated 1347 genes that are believed to be involved in breast cancer (Supplementary Table S6 of Hua *et al.*), referred to as the BC1000 genes. In all, 1156 of these genes remained after removing the probe sets that are always expressed at a low level and/or are without unique coordinates.

To determine whether EMBER was preferentially selecting the BC1000 genes as targets, we compared the number of genes that are potential targets (within 100 kb of a peak) to the number of genes selected as regulatory targets by EMBER. For example, for the RAR$\alpha$ binding sites, genes from the BC1000 list appear 2936 times as potential targets (many appear more than once, if there are multiple nearby binding sites) and EMBER selects 1131 of these as regulatory targets (38%). By comparison, there are 55 036 potential targets total, of which EMBER selects 10 660 as regulatory targets (19%). From these results, we calculate the overrepresentation of BC1000 genes in the EMBER targets as $\log_2(0.38/0.19)=1.0$. To assess significance, we define a binomial distribution by the number of BC1000 targets we would get if each had a probability of 0.19 of being selected as a target by EMBER. This distribution yields a $P$-value for the overrepresentation of BC1000 genes of $4.9\times10^{-127}$. In other words, EMBER shows a strong preference for the BC1000 genes. The overrepresentation of BC1000 genes for the 10 peak groups is shown in Figure 2; Supplementary Table S1 reports the $P$-values and targeting statistics by EMBER. Every peak set is overrepresented for BC1000 genes.

In contrast, we consider selecting targets using the more usual proximity method: the closest gene to each binding site (by distance to TSS) is chosen as the target. Repeating the above analysis for this method, we find essentially no preference for BC1000 genes in the targets, in stark contrast to EMBER. Given that these TFs are widely believed to be involved in the development of breast cancer, we expect the BC1000 genes to be targeted, as they are by EMBER but not by proximity alone.

We examine the relationship between EMBER and proximity in more detail in Section 3 in Supplementary Material and Supplementary Figure S4. In short, we find that a relationship with proximity arises in the EMBER targets, but this association is not strong enough to be predictive. Given that EMBER does appear to be predictive for the BC1000 genes, we can view EMBER as a method of limiting the false discovery rate (FDR) when looking beyond the closest genes for targets. Moreover, EMBER provides a well-defined framework for continuously adjusting the FDR by raising or lowering the score threshold for assignment as a target; we discuss the effects of threshold choice further in Section 3.5.

### 3.3 Verification of particular gene targets with EMBER

We also attempt to corroborate some of the key gene targets that were hypothesized previously. Hua *et al.* (2009) infer a simple regulatory network involving cross regulation of ER$\alpha$ with RAR$\alpha$ and RAR$\gamma$, and regulation of *FOXA1*, *FOS* and *GATA3* by these TFs. To see if we could recapitulate these interactions, we looked for these genes in the potential targets and EMBER targets for each peak group. Targeting by different peak groups, along with gene scores from EMBER, are given in Supplementary Table S2.

EMBER confirms targeting of *FOXA1*, *FOS* and *GATA3* by ER$\alpha$, RAR$\alpha$ and RAR$\gamma$. Examination of which genes are targeted by which sets of TFs, as detailed in Supplementary Table S2, is also illustrative. For instance, RAR$\alpha$ peaks that target *FOXA1* are almost all coincident with RAR$\gamma$, but there are several RAR$\gamma$ peaks not coincident with RAR$\alpha$, suggesting that RAR regulation of this gene is primarily through RAR$\gamma$. In contrast, the ER$\alpha$ sites are split: about half of those associated with *FOXA1*, *FOS* or *GATA3* are coincident with both RAR$\alpha$ and RAR$\gamma$, and half are not.

We also recapitulate the regulation of *ESR1* by RAR$\alpha$ and RAR$\gamma$. In this case, the regulation appears to be more RAR$\alpha$-dependent: all RAR$\gamma$ peaks that target *ESR1* are coincident with an RAR$\alpha$ peak. Interestingly, EMBER does *not* predict that ER$\alpha$ regulates the RARs. Although there are ER$\alpha$ binding sites around these genes, neither is chosen as a target. There are also RAR$\alpha$ and RAR$\gamma$ sites within 100 kb of their genes, from which one could infer some auto-regulation, but the EMBER results again do not support this targeting. While auto-regulation of *RARA* and *RARG*, or their regulation by ER$\alpha$, may occur in different contexts, from these data EMBER selects two insulin-like growth factor binding protein (IGFBP) genes as targets (*IGFBP4* instead of *RARA* and *IGBFP6* instead of *RARG*). Both these genes are on the BC1000 list.

### 3.4 Combinatorial effects of ER$\alpha$ and RAR$\alpha$

As mentioned in Section 2, EMBER produces not only gene targets, but also a log-odds score matrix, $S_{l,m}$, that can be used to understand how targeted genes are behaving. In particular, it can be useful to see how expression patterns differ for targets of different TF combinations. For example, although peak groups [e] and [$\alpha$] are disjoint by definition, and so their list of target genes differs significantly, their top expression patterns are quite similar. To quantify the similarity between two expression patterns $S^1_{l,m}$ and $S^2_{l,m}$, we define their distance as

$$d(S^1,S^2)=\frac{\sum_{l,m}\left|S^1_{l,m}-S^2_{l,m}\right|}{\sum_{l,m}\left|S^1_{l,m}\right|+\left|S^2_{l,m}\right|}. \qquad (2)$$

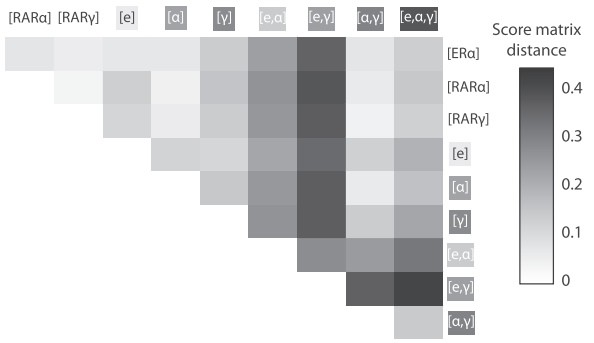


**Fig. 3.** Distance matrix between different TF binding site group score matrices. Distances are defined by Equation (2), and are between 0 and 1 by construction.

We include the denominator to normalize the sum, such that the distance is always between 0 and 1. A distance of 1 is achieved if the two score matrices differ in sign for every $(l,m)$ (a distance of 0 means that they are identical).

We plot the distances between the top expression patterns of the different peak groups in Figure 3. For the most part, the top expression patterns are quite similar: most distances are <0.15. However, groups [e,$\gamma$] (ER$\alpha$ and RAR$\gamma$, not RAR$\alpha$) and [e,$\alpha$] in particular stand out from the others. This suggests that coincident binding of ER$\alpha$ and either RAR$\alpha$ or RAR$\gamma$ lead to distinct modes of regulation. Hua *et al.* (2009) previously observed some differences in regulatory effects surrounding such coincident sites, such as the effect of RA and E2 co-treatment: looking at the nine putative target genes (from a proximity analysis, Supplementary Fig. S6 of above reference), they observe that the treatments had an antagonistic effect on co-targeted genes, but not on uniquely targeted genes. They also observed competition for some binding sites between ER$\alpha$ and RAR$\gamma$. Our analysis can provide more details about how those differences manifest. We focus on the combinatorial effects of ER$\alpha$ with RAR$\alpha$ because there are fewer peaks in [e,$\gamma$], so the differences in the elucidated behaviors of [e,$\gamma$] are more likely due to chance (Section 4 in Supplementary Material and Supplementary Fig. S6).

We use the score matrix for group [ER$\alpha$] as a reference for comparison with the score matrix for group [e,$\alpha$]. A visualization is presented in Figure 4. In analogy to the usual sequence logo representation of TF binding motifs, color bars indicating behaviors are scaled by how overrepresented the gene expression classification is, and then each behavior dimension is scaled by its relative entropy. The relative entropy (*RE*) for a given behavior dimension ($l$) is defined by

$$RE_l=\sum_{m=--}^{++} f_{1,l,m}\log(f_{1,l,m}/f_{2,l,m}) \qquad (3)$$

and quantifies how much of the information in the expression pattern distribution ($f_{1,l,m}$) can be explained by the background distribution ($f_{2,l,m}$). Thus, in Figure 4, large bars indicate highly overrepresented behaviors in highly information-rich behavior dimensions. Note that among the time course comparisons (first 15 behavior dimensions) in Figure 4, the most information-rich dimensions (tallest overall columns) are the early time point comparisons: 0–4 h for E2 treatment and 0 h (control) to 24 h for RA agonist treatment. Since the ChIP data were taken at the equivalent of the 0 h/control context,

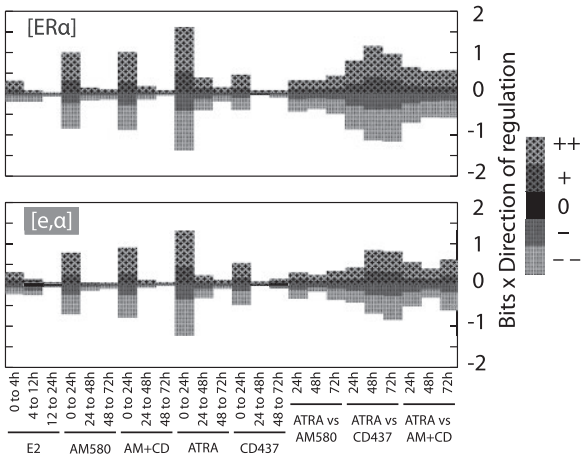


**Fig. 4.** Visualization of the score matrices for peak groups [ER$\alpha$] and [e,$\alpha$]. The values were first scaled to be between 0 and 1, then multiplied by the relative entropy. Down-regulated classifications ($--$ and $-$) are scaled below the horizontal axis, and up-regulated classifications ($++$ and $+$) are scaled above. The 0 classification is centered around the horizontal axis. Larger bars are over-represented classifications, and smaller bars are under-represented. For example, in the '0 to 24 h' comparison for ATRA treatment (10th behavior dimension from the left), classifications $++$ and $--$ are over-represented for groups [ER$\alpha$] and [e,$\alpha$], and classification 0 is the most under-represented (bar height is zero, so it does not appear). The behavior dimensions along the horizontal axis are in the same order as the list of comparisons in supplementary Fig. 2B.

it is not surprising that these early time comparisons contain the most useful information for distinguishing target genes.

In the [ER$\alpha$] score matrix, almost all the strongly up- and downregulated behaviors ($--$ and $++$) are overrepresented, and the no-change behaviors (0) are all underrepresented. The symmetry between up- and downregulated behaviors here is striking, which suggests that these binding sites are regulating genes in a mix of activation and repression. A heat map of the gene behaviors for the EMBER targets of [ER$\alpha$] is shown in Supplementary Figure S7, which provides further perspective on how this mix of up- and downregulation is manifested among the target genes.

The score matrix for group [e,$\alpha$] is qualitatively similar. Many of the aspects that make up the quantitative differences in the score matrix distance lend themselves to similar interpretations; for example, the $++$ behavior in E2: 0–4 h is overrepresented by a factor of 11 in [ER$\alpha$] but by a factor of 6.6 in [e,$\alpha$]. The more substantive changes regarding behavior are in some of the low *RE* dimensions. In particular, there is a loss of the symmetric behaviors in the second two behavior dimensions of E2 treatment (4–12 h and 12–24 h) and the AM580 treatment (24–48 h and 48–72 h). In each of these categories, the downregulated behaviors become dominant in the [e,$\alpha$] group. Since AM580 only ligates RAR$\alpha$, it is fitting that the differences appear in situations where ER$\alpha$ or RAR$\alpha$ are primarily activated. Finally, note that the expression patterns need not always be so similar. For example, in a recent study on mouse Pro-B cells (Mandal *et al.*, 2011), EMBER finds remarkably different expression patterns for STAT5 peaks coincident versus not coincident with H3k27me3 chromatin modifications (Fig. 7 of that reference).

**Table 1.** Pairwise comparison of top three expression patterns for all ER$\alpha$ peaks

| Expression pattern | No.of peaks | No. of target genes | $\log_2$ BC1000 bias ratio | *P*-value |
|---|---|---|---|---|
| 1 | 2493 | 3418 | 1.17 | $4.6\times10^{-60}$ |
| 2 | 3848 | 7292 | −0.78 | $4.0\times10^{-23}$ |
| 3 | 3148 | 4883 | −0.11 | $1.9\times10^{-1}$ |
| **Comparison** | **No. of same peaks** | **No. of same genes** | **Distance** | |
| 1 versus 2 | 1450 | 62 | 0.98 | |
| 2 versus 3 | 1905 | 60 | 0.94 | |
| 1 versus 3 | 1454 | 569 | 0.76 | |

Each expression pattern assigns target genes to some subset of the binding sites; remaining sites are considered 'non-regulatory' in the context of that expression pattern. For each expression pattern, we report the number of peaks and target genes (there are still some peaks with more than one gene, based on the standard threshold *T* used), and the $\log_2$ bias ratio for BC1000 genes, and the corresponding *P*-value. For reference, there are 9115 total ER$\alpha$ peaks, for which there are 19 864 potential targets. For each pair of expression patterns, we report the number of 'regulatory' peaks in common, the number gene targets shared and the distance between the score matrices.

### 3.5 Discovery of multiple regulatory modes

Like the DNA motif-finding algorithm MEME, EMBER can find multiple expression patterns. Less overrepresented patterns are found by using an erasing parameter as described in Section 1 of the Supplementary Materials, preventing genes that belong to a previously found expression pattern from contributing to the model for a subsequent expression pattern. However, it is still possible for a previously assigned gene to get re-assigned as a target if its behaviors match those of the new model by chance.

All the results described above use only the most overrepresented expression pattern for each group of TF binding sites. To illustrate the multiple expression pattern-finding property of EMBER, we obtained the top three expression patterns from the peaks in group [ER$\alpha$]. Comparisons between the expression patterns are shown in Table 1, and the expression patterns are shown in Supplementary Fig. S8. The second and third score matrices are distinctly asymmetric, in contrast to the balance of up- and downregulation noted in Section 3.4. The second score matrix is strongly biased toward upregulated behaviors (mostly red), and the third toward downregulated behaviors (mostly green).

A more quantitative analysis is provided in Table 1. We look first at the BC1000 log-bias ratio analysis discussed earlier for each expression pattern. As noted above, BC1000 genes are highly overrepresented in the targets for the first expression pattern. In contrast, they are highly underrepresented in the targets for the second expression pattern ($\log_2$ ratio is $-0.78$, and the *P*-value is $4.0\times10^{-23}$). In the third expression pattern, there is not any significant over- or underrepresentation. We also see the effect of the erasing parameter on the differences between expression patterns. Although the expression patterns share around half their peaks [indeed, some sites may regulate multiple genes (Bresnick and Felsenfeld, 1994; Pham *et al.*, 1996)], the assigned target genes are quite different (only $\sim$1% shared targets between expression Patterns 1 and 2, and expression Patterns 2 and 3), and the distance between the score matrices is close to the maximal value of 1.

Given the differences between these three expression patterns, the fact that any genes at all are shared may seem strange. However, shared genes between distinctly different expression patterns tend to be low scoring in both patterns. That is, they fit the models for the target genes only moderately well, but still better than they fit the background, so maximum likelihood comes from assigning those genes to the target set. To illustrate this effect, we raise the threshold [originally set at $T = \log(\lambda_2/\lambda_1)$] used to determine which genes are targets. Supplementary Table S4 gives the results when we multiply the threshold by a factor of 2, 4 or 6 and repeat the analysis in Table 1. Raising the threshold also results in fewer target genes per peak, so the overlap of peaks decreases as well.

As the threshold increases, the number of peaks and target genes associated with each expression pattern decreases, but the overlap decreases much faster. For example, 58% of the expression Pattern 1 peaks are shared with expression Pattern 2 at $1T$ in Table 1. This drops to 14% at a threshold of $6T$ (Supplementary Table S4). The number of genes in common drops to zero by $4T$. Re-examining the breast cancer data, we find that, as the threshold increases, the bias ratio becomes more exaggerated but the significance drops. For example, going from $1T$ to $6T$, the expression Pattern 1 bias ratio increases from 1.17 to 1.63 but the $P$-value increases (i.e. significance decreases) from $4.6 \times 10^{-60}$ to $3.0 \times 10^{-28}$.

The above examples suggest that the threshold can be used as a continuous control over how many genes are allowed to be considered targets. Presumably, raising the threshold corresponds to reducing the number of false positives at the expense of more false negatives. For the analyses presented in this work, we have used the standard threshold $T$ (which maximizes the likelihood) throughout, but it clearly may be beneficial for researchers to consider higher thresholds. The choice depends on how the results will be used. For example, if further statistical analysis is to be performed, it may be better to use a lower threshold to retain more genes. However, if the next step is experimental validation or testing of particular targets, then it makes sense to start with the highest scoring genes first. Part of the utility of EMBER is that it provides a quantitative, systematic framework for deciding which genes are the strongest targets.

Finally, it should be noted that not all expression patterns returned by EMBER necessarily represent real regulatory modes, just as not all DNA motifs discovered by an algorithm like MEME represent actual binding sites. Given the results for ER$\alpha$ described here, it is reasonable to believe that the second expression pattern represents a secondary regulatory mode of ER$\alpha$ that is not associated with breast cancer development. On the other hand, within our limited analysis there is not strong evidence that the third expression pattern is representative of an actual regulatory mode.

## 4 CONCLUSIONS

We have presented EMBER, a method for inferring gene regulatory targets of TF binding sites by integrating gene expression information from DNA microarray experiments. Using a hand-annotated set of genes involved in breast cancer, we demonstrated that EMBER implicates the TFs ER$\alpha$, RAR$\alpha$ and RAR$\gamma$ in breast cancer development, and that assigning targets by proximity does not. However, it is likely that further improvements could be made to the target selection method in EMBER. For instance, a behavior dimension selection step could be added to the method, as in some quantitative structure–activity relationship (QSAR) methods (So and Karplus, 1996). Alternatively, because gene behaviors will likely be highly correlated in different behavior dimensions, it may be helpful to consider joint distributions between behavior dimensions in EMBER. More importantly, recent advances in sequencing technology have greatly increased the amount and variety of data available to researchers, and many of these data sources are relevant to the problem of inferring TF regulation. Possible additional sources of data include ChIP-seq of histone modifications, which correspond to different types of regulatory activity (Barski *et al.*, 2007; Bernstein *et al.*, 2005; Birney *et al.*, 2007; Ernst *et al.*, 2011; Guenther *et al.*, 2007; Heintzman *et al.*, 2007, 2009; Mikkelsen *et al.*, 2007); chromosome conformation capture, which measures which segments of DNA are in close spatial contact with one another (Dekker *et al.*, 2002; Dostie and Dekker, 2007); and global run-on and sequencing (GRO-seq), which directly measures transcriptional activity genome-wide (Hah *et al.*, 2011). Although we have not made use of any of the above data types, they could be used pre-EMBER to curate the potential regulatory targets or post-EMBER to confirm the regulatory predictions. It may be possible to incorporate information from these methods into EMBER itself by using them analogously to either the TF binding sites or the expression levels. These possibilities will be investigated in the future.

*Funding*: Department of Energy Computational Science Graduate Fellowship program; Chicago NIH Systems Biology Center (P50 GM081892).

*Conflict of Interest*: none declared.

## REFERENCES

Arnosti,D. and Kulkarni,M. (2005) Transcriptional enhancers: intelligent enhanceosomes or flexible billboards? *J. Cell Biochem.*, **94**, 890–898.

Bailey,T. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA, pp. 28–36.

Banerji,J. *et al.* (1981) Expression of a $\beta$-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, **27**, 299–308.

Bar-Joseph,Z. *et al.* (2003) Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.*, **21**, 1337–1342.

Barski,A. *et al.* (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.

Berger,M. and Bulyk,M. (2006) Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins. *Methods Mol. Biol.*, **338**, 245–260.

Bernstein,B. *et al.* (2005) Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell*, **120**, 169–181.

Birney,E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

Boulesteix,A.-L. and Strimmer,K. (2005) Predicting transcription factor activities from combined analysis of microarray and ChIP data: a partial least squares approach. *Theor. Biol. Med. Model*, **2**, 23.

Bresnick,E. and Felsenfeld,G. (1994) Dual promoter activation by the human beta-globin locus control region. *Proc. Natl Acad. Sci. USA*, **91**, 1314–1317.

Capaldi,A. *et al.* (2008) Structure and function of a transcriptional network activated by the MAPK Hog1. *Nat. Genet.*, **40**, 1300–1306.

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.

Dekker,J. *et al.* (2002) Capturing chromosome conformation. *Science*, **295**, 1306–1311.

Dostie,J. and Dekker,J. (2007) Mapping networks of physical interactions between genomic elements using 5C technology. *Nat. Protoc.*, **2**, 988–1002.

Eisen,M. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Ernst,J. *et al*. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–51.

Fujiwara,T. *et al*. (2009) Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. *Mol. Cell*, **36**, 667–681.

Gao,F. *et al.* (2004) Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics*, **5**, 31.

Gilchrist,M. *et al*. (2006) Systems biology approaches identify ATF3 as a negative regulator of Toll-like receptor 4. *Nature*, **441**, 173–178.

Guenther,M. *et al.* (2007) A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, **130**, 77–88.

Hah,N. *et al.* (2011) A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell*, **145**, 622–634.

Heintzman,N.D. *et al*. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, **39**, 311–318.

Heintzman,N.D. *et al*. (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**, 108–112.

Hua,S. *et al*. (2008) Genomic analysis of estrogen cascade reveals histone variant H2A.Z associated with breast cancer progression. *Mol. Syst. Biol.*, **4**, 188.

Hua,S. *et al.* (2009) Genomic antagonism between retinoic acid and estrogen signaling in breast cancer. *Cell*, **137**, 1259–1271.

Johnson,D. *et al.* (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.

Kim,H. *et al.* (2009) Transcriptional regulatory circuits: predicting numbers from alphabets. *Science*, **325**, 426–432.

Lee,G. *et al.* (2005) Hypersensitive site 7 of the TH2 locus control region is essential for expressing TH2 cytokine genes and for long-range intrachromosomal interactions. *Nat. Immunol.*, **6**, 42–48.

Li,C. and Wong,W. (2003) DNA-Chip analyzer (dChip). In Gail,M. *et al.* (eds) *The Analysis of Gene Expression Data.* Statistics for Biology and Health. Springer, London, pp. 120–141.

Liu,R. *et al*. (1998) Interaction of BAG-1 with retinoic acid receptor and its inhibition of retinoic acid-induced apoptosis in cancer cells. *J. Biol. Chem*., **273**, 16985–16992.

MacQueen,J. (1967) Some methods for classification and analysis of multivariate observations. In *Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1*. University of California Press, California, pp. 281–297.

Mandal,M. *et al*. (2011) Epigenetic repression of the *Igk* locus by STAT5-mediated recruitment of the histone methyltransferase Ezh2. *Nat. Immunol.*, **12**, 1212–1220.

McLean,C. *et al*. (2010) GREAT improves functional interpretation of *cis*-regulatory regions. *Nat. Biotechnol.*, **28**, 495–503.

Mikkelsen,T.S. *et al*. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.

Pepke,S. *et al.* (2009) Computation for ChIP-seq and RNA-seq studies. *Nat. Methods* , **6**, S22–S32.

Pham,C. *et al.* (1996) Long-range disruption of gene expression by a selectable marker cassette. *Proc. Natl Acad. Sci. USA*, **93**, 13090–13095.

Ptashne,M. (1992) *A Genetic Switch: Phage λ and Higher Organisms*. Blackwell Science & Cell Press, Cambridge, MA.

Robertson,G. *et al*. (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, **4**, 651–657.

Sanguinetti,G. *et al.* (2003) Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities. *Nat. Biotechnol.*, **21**, 1337–1342.

Sanguinetti,G. *et al.* (2006) A probabilistic dynamical model for quantitative inference of the regulatory mechanism of transcription. *Bioinformatics*, **22**, 1753–1759.

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

So,S.-S. and Karplus,M. (1996) Evolutionary optimization in quantitative structure-activity relationship: an application of genetic neural networks. *J. Med. Chem.*, **39**, 1252–1530.

Townsend,J. and Hartl,D. (2002) Bayesian analysis of gene expression levels: statistical quantification of relative mRNA level across multiple strains or treatments. *Genome Biol.*, **3**, RESEARCH0071.

Valouev,A. *et al*. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-seq data. *Nat. Methods*, **5**, 829–834.

Verzi,M. *et al*. (2010) Differentiation-specific histone modifications reveal dynamic chromatin interactions and partners for the intestinal transcription factor CDX2. *Dev. Cell*, **19**, 713–726.

Wu,H. *et al.* (2003) MAANOVA: a software package for the analysis of spotted cDNA microarray experiments. In Gail,M. *et al.* (eds) *The Analysis of Gene Expression Data*, Statistics for Biology and Health. Springer, London, pp. 120–141.

Yager,J. and Davidson,N. (2006) Estrogen carcinogenesis in breast cancer. *N. Engl. J. Med.*, **354**, 270–282.

Yu,M. *et al*. (2009) Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis. *Mol. Cell*, **36**, 682–695.

Zhou,Q. *et al.* (1997) Inhibition of cyclin D expression in human breast cancer carcinoma cells by retinoids in vitro. *Oncogene*, **15**, 107–115.