



**HAL**  
open science

## Computational discovery of regulatory elements in a continuous expression space

Mathieu Lajoie, Olivier Gascuel, Vincent Lefort, Laurent Brehelin

► **To cite this version:**

Mathieu Lajoie, Olivier Gascuel, Vincent Lefort, Laurent Brehelin. Computational discovery of regulatory elements in a continuous expression space. *Genome Biology*, 2012, 13 (11), pp.41. 10.1186/gb-2012-13-11-r109 . lirmm-00796453

**HAL Id: lirmm-00796453**

**<https://hal-lirmm.ccsd.cnrs.fr/lirmm-00796453>**

Submitted on 4 Mar 2013

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

This Provisional PDF corresponds to the article as it appeared upon acceptance. Copyedited and fully formatted PDF and full text (HTML) versions will be made available soon.

## Computational discovery of regulatory elements in a continuous expression space

*Genome Biology* 2012, **13**:R109 doi:10.1186/gb-2012-13-11-r109

Mathieu Lajoie (mathieu.lajoie@gmail.com)

Olivier Gascuel (gascuel@lirmm.fr)

Vincent Lefort (lefort@lirmm.fr)

Laurent Brehelin (brehelin@lirmm.fr)

**ISSN** 1465-6906

**Article type** Method

**Submission date** 4 June 2012

**Acceptance date** 25 October 2012

**Publication date** 27 November 2012

**Article URL** <http://genomebiology.com/2012/13/11/R109>

This peer-reviewed article can be downloaded, printed and distributed freely for any purposes (see copyright notice below).

Articles in *Genome Biology* are listed in PubMed and archived at PubMed Central.

For information about publishing your research in *Genome Biology* go to

<http://genomebiology.com/authors/instructions/>

# Computational discovery of regulatory elements in a continuous expression space

Mathieu Lajoie\* , Olivier Gascuel , Vincent Lefort , Laurent Bréhélin

Méthodes et algorithmes pour la Bioinformatique, LIRMM, Univ. Montpellier 2, CNRS; 161 rue Ada, 34095 MONTPELLIER, France

Email: Mathieu Lajoie - mathieu.lajoie@gmail.com; Olivier Gascuel - gascuel@lirmm.fr; Vincent Lefort - lefort@lirmm.fr; Laurent Bréhélin - brehelin@lirmm.fr;

\*Corresponding author

## Abstract

---

Approaches for regulatory element discovery from gene expression data usually rely on clustering algorithms to partition the data into clusters of co-expressed genes. Gene regulatory sequences are then mined to find over-represented motifs in each cluster. However, this ad-hoc partition rarely fits the biological reality. We propose a novel method called RED2 that avoids data clustering by estimating motif densities locally around each gene. We show that RED2 detects numerous motifs not detected by clustering-based approaches, and that most of these correspond to characterized motifs. RED2 can be accessed online with a user-friendly interface at <http://www.atgc-montpellier.fr/RED2/>.

---

**Keywords:** motif discovery; gene regulation; co-expressed genes; clustering; k-nearest neighbors

## Background

Gene expression is modulated depending on time, space and environmental conditions. This process involves many levels, including chromatin structure, transcription, transcript stability or localization, and translation. At most of these levels, regulation is achieved through the binding of regulatory proteins with specific short nucleic acid sequences called regulatory elements (RE). A notable exception is

ncRNA-mediated regulation [1].

Discovering RE (*i.e.* their associated sequence motifs) from gene expression and sequence data can be done in several ways. One of the most common approaches is to use a clustering algorithm to partition the expression dataset, and to apply, on each cluster, one of the numerous algorithms that have been designed to find over-represented motifs in a predefined set of sequences, such as MEME [2], AlignACE [3] or Weeder [4] (see Sandve *et al.* [5] for a comparison of their performance). There are two major limitations to this approach:

- First, most of these algorithms rely on statistical models of sequence background, which have been reported to produce many false positives [6, 7], especially with repeat-rich and atypical genomes. For example, this is the case with *Plasmodium falciparum*—the main causal agent of human malaria—whose A+T content reaches almost 90% in intergenic regions [8].
- Second, clustering assumes that i) a “natural” partition of the genes in well-defined clusters exists and can be inferred from the data, and ii) co-regulated gene sets are disjoint. In fact, determining the “real” number of clusters in a dataset is considered to be one of the hardest classification problems, and has been the topic of numerous studies. Moreover, it is a well known fact that different regulatory motifs can have overlapping gene sets [9, 10]. These two points are illustrated in Figure 1 with the *P. falciparum* intraerythrocytic gene expression data of Bozdech *et al.* [11]. The periodic nature of this expression dataset allows projection on a two dimensional space with minimal loss of variance using principal component analysis. The upper part of the figure represents the partitions obtained after performing k-means clustering of the genes in the original space using different numbers of clusters. We see that the clusters have similar shapes and sizes and that their edges arbitrarily change depending on the number of clusters. The resulting beach ball-like partitions of the expression space suggest that there is no natural clustering in this dataset. In the bottom part of the figure, we show the regions of the expression space enriched for three regulatory motifs identified in previous studies [6, 7]. We see that the regions defined for the different motifs clearly overlap and do not correspond to the regions defined by any of the above clusterings.

Several methods that do not require a clustering have also been developed. For example, RankMotif++ [12] and DRIM [13] detect motif imbalance in ranked lists of genes obtained by ordering expression changes between two conditions. These methods have the advantage of not relying on a statistical model of background sequence, as they consider the motif distribution in the whole set of

available sequences. However, although well-suited for two-condition datasets, a single ordering can hardly reflect all the information contained in high dimensional datasets, for example those obtained by measuring expression in multiple conditions or at different time-points, as routinely produced nowadays. Similarly, methods like REDUCE [14] and MatrixREDUCE [15] do not rely on clustering. They use linear regression to identify motifs that can be used to predict expression levels, assuming that the latter are linearly correlated with the number of motif occurrences in each sequence. However, as noted in Elemento *et al.* [6], the validity of this assumption has not been widely explored, nor the behavior of such approaches in case of violation. Finally, we mention the early work of Holmes and Bruno [16], which discovers the clustering and the motifs simultaneously. Although this approach seems very appealing, it has not been tested on real datasets, probably because of its prohibitive running time [16].

FIRE [6] and GEMS [7] algorithms have been specifically designed for finding RE from whole genomes and high dimensional datasets. They differ advantageously from DRIM and RankMotif++ as they are not restricted to ranked lists of genes, and from REDUCE and MatrixREDUCE as they make as few assumptions as possible about the way in which RE drive expression. Compared to MEME and AlignACE, they do not use a background model and are less prone to reporting false positives. However, they both rely on expression-data clustering, and thus are subject to the criticisms mentioned above. GEMS differs from FIRE in the way the dependency between the presence of a motif and the expression profile of the corresponding gene is measured. Specifically, GEMS uses the hypergeometric distribution to assess motif enrichment in each cluster, while FIRE computes the mutual information between the presence/absence of a motif and the cluster membership of the corresponding sequences. These two methods can be seen as two extremes of a simple approach that assumes the RE distribution must show some kind of statistical dependency with the expression data. The hypergeometric approach is a local criterion, as it considers motif enrichment in a single co-expression cluster, while the mutual information approach is a global one, as it includes contributions from all clusters. Since they make few assumptions, these two methods can be applied to a wide range of expression data, and they are particularly useful when knowledge of the underlying regulatory mechanisms is limited. FIRE and GEMS accurately predicted several of the *P. falciparum* ApiAP2 binding sites (experimentally validated in [17,18]), and FIRE has recently been used to detect RNA recognition elements in *Saccharomyces cerevisiae* [19].

In this work, we show how the hypergeometric distribution (local criterion) and the mutual information (global criterion) can be used without requiring any clustering, using the notion of *motif density* in the

expression space. Rather than considering the number of genes that contain a motif in each co-expression cluster, we estimate motif densities locally around each gene by considering the proportion of neighboring genes (in the expression space) that contain the motif. The remainder of the paper is organized as follows. We first present the continuous versions of the hypergeometric and mutual information criteria, and the RED<sup>2</sup> software that implements them. Second, we describe the output of RED<sup>2</sup> and explain how the motif density can be used to assess the functionality of a particular motif occurrence. Third, we compare the original and new versions of the two criteria on several expression datasets from yeast, and show that using the new versions usually results in a significant increase in the motif detection sensitivity. In addition, we perform a similar comparison with two well-established methods, FIRE and MatrixREDUCE, which demonstrates the broad utility of our approach. Finally, as a case study, we apply RED<sup>2</sup> on two *P. falciparum* datasets and discuss some findings that can be deduced from the results.

## Results and discussion

We are given a set  $G$  of  $n$  genes with two kinds of data. First, each gene is associated with a vector that describes its expression level at different time points or in a given set of conditions (*i.e.* its expression profile). Second, each gene is associated with a nucleic acid sequence, which may correspond to its upstream, intronic or downstream region. In addition, we are given a function to compute the distance between the expression profile of two genes—generally, this is the Euclidean or the Pearson correlation distance. As noted in the introduction, the goal is to identify motifs whose presence or absence in the sequences shows statistical dependency with the expression data.

Broadly speaking, a motif  $m$  is a set of sequences defined over  $\{A,C,G,T\}$ . We say that a sequence  $s$  contains a motif  $m$  if at least one substring of  $s$  belongs to  $m$ . Two popular ways of representing motifs are the Position Weight Matrices (PWMs) and consensus strings with degenerate symbols (IUPAC system) [5]. Here we use the IUPAC system, representing motifs with special characters encoding alternative nucleic acids—for example, the WAGACA motif corresponds to  $\{AAGACA,TAGACA\}$ . In theory, PWMs are more expressive than IUPAC strings, as they can represent more motifs and associate a score to each possible occurrence. However, PWMs require a score threshold to determine whether a particular sequence is a match or not, which may be hard to set. Moreover, they have been shown to perform similarly to IUPAC strings in a benchmark study [5], and, from a computational perspective, dealing with IUPAC strings is less demanding than PWMs.

## Scoring functions

Here we describe the local and global criteria used by GEMS and FIRE, respectively, and show how they can be used without clustering, by computing motif density locally around each gene in the expression space. Namely, given a gene  $g \in G$  and a motif  $m$ , the motif density around  $g$  is simply the proportion of genes that possess  $m$  among the neighbors of  $g$ . This involves defining a neighborhood for each gene in the expression space. We could use the  $k$ -nearest neighbors of the gene according to the chosen distance function (Euclidean or Pearson), but we will see that using this neighborhood may reduce the sensitivity of the approach because of the *hubness* phenomenon [20]. Therefore, after presenting the two criteria, we propose an alternative solution to define the neighborhood.

### Local criterion

The GEMS algorithm expresses the dependency between a motif and the expression data in terms of motif enrichment, assuming that a motif is likely to have a biological function if, in any given co-expression cluster, it is present in more genes than would be expected by chance. Let  $c$  be a co-expression cluster containing  $n_c$  genes and let  $n_{mc}$  denote the number of genes having the motif  $m$  in  $c$ . Under the null hypothesis that  $m$  is distributed independently of the expression data, the probability of observing  $X \geq n_{mc}$  positive genes (*i.e.* containing  $m$ ) in  $c$  is derived from the hypergeometric distribution:

$$P(X \geq n_{mc}) = \sum_{j=n_{mc}}^{\min(n_c, n_m)} \frac{\binom{n_m}{j} \binom{n-n_m}{n_c-j}}{\binom{n}{n_c}},$$

where  $n_m$  is the total number of genes with motif  $m$  in  $G$ . For a given clustering  $\mathcal{C}$ , the motif enrichment score of  $m$  is defined in GEMS as

$$s(m, \mathcal{C}) = \max_{c \in \mathcal{C}} (-\log P(X \geq n_{mc})). \quad (1)$$

This is a local criterion because the score of the motif is given by the single cluster where the enrichment is the most significant.

A straightforward way of modifying this function for non-clustered data is to take each gene  $g \in G$  in turn, and to consider the number of positive genes (*i.e.* which have the motif) in its neighborhood. Let  $n_{mk_g}$  denote the number of positive genes among the  $k$  neighbors of  $g$ . This leads to

$$s(m, G) = \max_{g \in G} (-\log P(X \geq n_{mk_g})). \quad (2)$$

A major difference between Equations (1) and (2) is that the latter considers different sets of overlapping genes, which is biologically more sound than considering a single partition when seeking multiple sets of

co-regulated genes. However, an obvious issue with Equation (2) is the multiplication of tests required for each motif. We will see that this is not a problem in practice because this quantity is not utilized as a p-value but rather as a score function, which is used to compute a false discovery rate.

### Global criterion

The FIRE algorithm expresses the dependency between a motif  $m$  and the co-expression clusters in terms of mutual information. Assume that we randomly pick a gene  $g \in G$  and let  $C$  be a random variable whose outcome is the cluster membership of  $g$ . Similarly, let  $M$  be a random variable whose outcome is 1 if  $g$  has motif  $m$ , and 0 otherwise. The mutual information between  $M$  and  $C$  is defined in FIRE as

$$I(M; C) = \sum_{M, C} P(M, C) \log \frac{P(M, C)}{P(M)P(C)}, \quad (3)$$

with  $P(C) = n_c/n$ . Moreover,  $P(M)$  and  $P(M, C)$  are written as

$$P(M) = \begin{cases} \frac{n_m}{n}, & \text{if } M = 1 \\ 1 - \frac{n_m}{n}, & \text{if } M = 0 \end{cases}$$

and

$$P(M, C) = \begin{cases} \frac{n_{mc}}{n}, & \text{if } M = 1 \\ 1 - \frac{n_{mc}}{n}, & \text{if } M = 0. \end{cases}$$

To get rid of the clustering, let us consider the probability density function over the space of the gene expression profiles. The gene expression profiles are assumed to be sampled from this density. The mutual information between the presence/absence of a motif  $m$  and a gene expression profile  $X$  is equal to

$$\begin{aligned} I(M; X) &= \int_X \sum_M P(M, X) \log \frac{P(M, X)}{P(M)P(X)} dX \\ &= \int_X P(X) \left( \sum_M P(M|X) \log \frac{P(M|X)}{P(M)} \right) dX. \end{aligned}$$

This expression can be interpreted as the expected value of the function  $f(X) = \sum_M P(M|X) \log \frac{P(M|X)}{P(M)}$  with respect to the profile density function. Hence,  $I(M; X)$  can be estimated from the average

$$\frac{1}{n} \sum_{g \in G} \sum_M P(M|X_g) \log \frac{P(M|X_g)}{P(M)}, \quad (4)$$

where  $X_g$  is the expression profile of gene  $g$ . Probability  $P(M|X_g)$  is estimated using the motif density around  $g$ , *i.e.* the proportion of positive genes among the  $k$  neighbors of  $g$ :

$$P(M|X_g) = \begin{cases} \frac{n_{mk_g}}{k}, & \text{if } M = 1 \\ 1 - \frac{n_{mk_g}}{k}, & \text{if } M = 0. \end{cases}$$

### *Neighborhood definition*

In the above formulae (Equations (2) and (4)), we use the number of genes that contain the motif among the  $k$  neighbors of each gene. A natural solution to define this neighborhood is to consider the  $k$ -nearest neighbors (kNN) of the gene. However, a major drawback of this approach stems from the *hubness* phenomenon [20] which occurs in high-dimensional space. This phenomenon can be described as the tendency for a few points (genes) to appear in an excessive number of neighborhoods, while many other genes appear in zero or few neighborhoods. If we encode the kNN relationships into a directed graph by putting an arc from node (gene)  $g$  to node  $g'$  if  $g'$  is among the  $k$ -nearest neighbors of  $g$ , we observe that the in-degree distribution of this graph—*i.e.* the number of arcs ending on a node—is strongly skewed to the left. This is exemplified in the left panel of Figure 2, which is obtained with the Gasch et al. [21] dataset, a compendium of yeast transcriptomic measurements in a variety of stress conditions (see experiments below). Such a neighborhood biases our overall density estimation, since it gives more weight to the most popular neighbors, while discarding the information provided by the least popular ones. Radovanovic et al. [20] showed that this phenomenon can affect both supervised (e.g. kNN predictor) and unsupervised (e.g. k-means) classification algorithms that use neighborhoods. In our case, the effect is particularly important, as it can prevent the discovery of motifs associated with genes that do not belong to the hubs. To circumvent this problem, we have designed a procedure that starts from the kNN graph described above. In this graph, the neighborhood of a gene  $g$  involves all genes  $g'$  such that there is an arc  $g \rightarrow g'$ . First, the neighborhood relations are symmetrized: if there is an arc  $g \rightarrow g'$  and no arc  $g' \rightarrow g$ , the latter is added to the graph. In this way, each gene is in the neighborhood of at least  $k$  other genes. We then use this symmetrized graph to construct a second graph in the following way. For each gene, we select  $k$  neighbors by randomly sampling without replacement among its neighborhood, using a probability distribution inversely proportional to the number of neighborhoods in which the neighbors appear (see Material and Methods for details). Hence, genes that are in many neighborhoods have less chance to appear in each of these neighborhoods. As we observe in the right panel of Figure 2, this strategy reduces the asymmetry of the in-degree distribution of the neighborhood graph. It would likely be possible to define other procedures to achieve this goal, but, as we will see in the experiments, this one effectively increases the sensitivity of our motif detection approach.

## The RED<sup>2</sup> software

The new versions of the scoring functions have been implemented (C++) in a software called RED<sup>2</sup> (for *Regulatory Element Discovery from Raw Expression Data*), which can be run from <http://www.atgc-montpellier.fr/RED2/>. RED<sup>2</sup> takes as input a set of expression profiles and their associated DNA sequences, and outputs a set of IUPAC motifs with various statistics, such as their distribution in the expression space and in the input sequences. To make the analyses easier, the promoter sequences of several model organisms are pre-loaded on the web server. Like many methods for motif discovery, RED<sup>2</sup> starts by computing the score of every possible  $q$ -mer ( $q = 7$  by default) and then proceeds with a local optimization of the highest scoring  $q$ -mers (called the seeds), transforming them into richer and better scoring motifs. When all seeds have been optimized, a filtering step removes the redundancy in the set of inferred motifs. We describe RED<sup>2</sup> implementation and output below.

### *Seed identification step*

To assess the significance of a seed score while taking into account multiple testing, we estimate the proportion of false positives (False Discovery Rate,  $FDR$ ) for different score thresholds. To this end, we rank  $q$ -mers according to their score and compare the resulting list with the expected score distribution under hypothesis  $H_0$  that there is no dependency between  $q$ -mers and expression profiles. This distribution is estimated by randomly shuffling the mapping between the nucleic acid sequences and the expression profiles, and re-evaluating all  $q$ -mer scores several times (10 by default). The  $FDR$  associated with an observed score  $s$  is estimated with the formula

$$\widehat{FDR}(s) = \frac{\text{expected number of } q\text{-mers with a score } \geq s \text{ under } H_0}{\text{number of } q\text{-mers with score } \geq s \text{ in original data}}. \quad (5)$$

$q$ -mers with score associated with an  $FDR$  below a given threshold (0.001 by default) are then considered for optimization. Note that the above formula may sometimes be  $> 1$ . In this case the  $FDR$  is assumed to be 1 for the considered score.

### *Optimization step*

Each seed is iteratively refined in a greedy way until its score reaches a local maximum. At each iteration, we evaluate the score of all solutions (motifs) obtained by replacing any non-degenerate position of the current solution by any IUPAC symbol generalizing the nucleotide at this position, and the best scoring solution is selected for the next iteration. For example, the motif WAGACA will lead to the set

$\{WWGACA, WHGACA, \dots, WARACA, \dots\}$ , where  $W = \{A, T\}$ ,  $H = \{A, T, C\}$  and  $R = \{A, G\}$ .

A critical point when generalizing the seeds is to avoid constructing a motif whose sequence description looks similar to that of the seed, but which is present in genes associated with very different expression profiles. To prevent such “incoherent” motifs, RED<sup>2</sup> considers a motif  $m$  that generalizes a seed  $s$  as a potential solution only if the genes covered by  $m$  are close to those covered by  $s$  in the expression space. To this end, RED<sup>2</sup> uses the density profiles of the motifs, *i.e.* the vector of length  $n$  which describes the density of the motif around each gene in  $G$ . Recall that the density of  $m$  around a gene  $g$  is the proportion of genes with the motif in the neighborhood of  $g$ , *i.e.*  $\frac{n_{mk_g}}{k}$ . Then,  $m$  is considered as a potential optimization of  $s$  only if the Pearson correlation between the density profiles of  $m$  and  $s$  is above a fixed threshold  $\alpha$  ( $\alpha = 0.75$  by default). This point is illustrated in Figure 3, which compares the density profiles of two pairs of close 6-mers in *P. falciparum* using the Bozdech dataset (see the application to *P. falciparum* below). The 6-mers of the first pair have similar density profiles, suggesting that they are part of the same regulatory motif. On the contrary, the 6-mers of the second pair have quite different profiles, suggesting that they are bound by different regulatory proteins.

When this first optimization step has reached a local maximum, RED<sup>2</sup> tries to elongate each motif with a similar procedure. In this second step, solutions are obtained by adding any IUPAC symbol to one of the motif extremities. The parameter  $\alpha$  is not used here, as this step can only decrease the set of sequences in  $m$ . The maximum motif length is set at 9 by default, but it can be increased up to 15.

### *Filtering step*

Several seeds may be associated with the same biological motif, and the optimization step usually leads to a highly redundant set of motifs. To retain the best variant of each motif, RED<sup>2</sup> compares them in two ways. First, it computes the overlap between each pair of motifs. The overlap is defined as the maximum number of positions that can be aligned without internal gaps or mismatches (two IUPAC symbols *match* if their intersection is non-empty). For example, the overlap between WAGACA and TATAGA is 4, corresponding to the alignment

-	-	W	A	G	A	C	A
T	A	T	A	G	A	-	-

By default, RED<sup>2</sup> considers that two motifs may be redundant if their overlap is at least 4. However, as discussed above, similar motifs (at the sequence level) may be associated with different locations in the expression space, and hence correspond to different transcription factors. Therefore, to be considered

redundant, the Pearson correlation between their density profiles must also be above a fixed threshold  $\gamma$  ( $\gamma = 0.75$  by default).

The filtering step proceeds as follows. All optimized seeds are sorted according to their score and the best scoring one is put in the (empty) set of filtered motifs. Then the remaining ones are considered sequentially and added to the set of filtered motifs if they are not redundant with respect to that set.

### **RED<sup>2</sup> output and functionality of motif occurrences**

For each motif, RED<sup>2</sup> returns the logo of the motif together with various statistics and graphics (see Figure 4). Firstly, a histogram describes the distribution of motif occurrence positions relative to the 5' or 3' sequence extremities (see middle of Figure 4). Secondly, a heatmap (right of Figure 4) shows the distribution of the motif in the expression space. The x-axis of this map represents the different conditions in the expression dataset, while the y-axis represents the expression level. The y-axis is divided into 25 bins, and the colors of the map indicate the motif density in each of these bins, expressed as a z-score (red for high and green for low ; see motif density in Material and methods). For example, in Figure 4 which shows three motifs identified in *P. falciparum* from the Bozdech dataset (see the application to *P. falciparum* below), we can see that the first motif (Fig. 4(a)) is over-represented in genes that have low expression levels in the first third of the conditions (*i.e.* in the first 16 hours of the erythrocytic cycle), high expression levels in the following hours, and again low levels in the last few hours. On the contrary, the third motif (Figure 4(c)) is over-represented in genes that have quite opposite profiles.

For each motif, RED<sup>2</sup> uses a sign test to assess a potential strand bias (*i.e.* if the motif occurs preferentially on the forward or reverse strand), and provides a GO enrichment analysis of the genes possessing the motif. Moreover, it outputs the position and context (10 bp on each side) of all occurrences, and the list of genes that possess the motif, sorted by motif density (z-score). This sorted list is of high practical interest, because the motif density around a gene where a particular occurrence is found can be a good indicator of that occurrence's functionality. The rationale is that a motif occurrence is more likely to be functional in a gene that is in a region of the expression space where many genes also possess the motif. This is illustrated in Figure 5 with the PAC/TOD6 motif, whose functionality is highly constrained by its position in the *S. cerevisiae* promoter sequences [22]. In this figure, the position of all PAC motif occurrences are plotted against their local density with respect to the Gasch et al. dataset (see the experiments below). We can see a strong positional bias for occurrences in high motif density regions, while occurrences in low density regions are more uniformly distributed. Similar results are obtained with the SFP1 and RPN4 motifs.

This discriminative feature can be used to refine the selection of genes that are likely regulated by a given motif. To account for the prior probability of each motif, and to better distinguish between high and low density values, RED<sup>2</sup> actually expresses the densities as a z-score, which is the difference between the fraction of genes that possess the motifs and the expected value under the null hypothesis, in terms of standard deviation (see Formula 6 in Material and methods). Thus, positive z-scores indicate genes in regions of the expression space where the proportion of genes that have the motif is higher than expected, while negative z-scores indicate lower than expected proportions.

### **Assessment on yeast**

We assessed our approach on yeast, which is likely the best known eukaryote for transcriptional regulation. We used 24 expression datasets: 22 datasets from the Gasch et al. [21] compendium, the complete compendium itself and the Spellman et al. [23] dataset (see Material and Methods). The 600 bp upstream of each gene was used in these experiments.

### *Hubness phenomenon*

First, we assessed the effect of the hubness phenomenon on the new version of the scoring functions. We mentioned previously that the  $k$ -nearest neighborhood is subject to this phenomenon, and we proposed an alternative neighborhood based on the *symmetric*  $k$ -nearest neighbors. To measure how this affects the performance of our approach, we compared the number of 7-mers detected at different  $FDR$  thresholds using both types of neighborhood. Figure 6 shows the results obtained with a neighborhood of size 200 on the Gasch et al. dataset [21] (complete compendium). As we can see in this figure, the hubness phenomenon can have a huge effect, especially for the mutual information scoring function, and a proper definition of the gene neighborhood is crucial to achieving good results.

### *Seed identification*

We compared the original and continuous versions of the scoring functions on the 24 yeast expression datasets. The clustering used for the original scoring functions was obtained with the k-means algorithm of the R project [24]. For each dataset, we tried several cluster numbers: 5, 10, 20, 40, 80. For each number, the k-means algorithm was run 100 times and the best clustering (according to the k-means criterion) was retained for the experiments below. For the continuous version of the scoring functions (Equations (2) and (4)), we always used a neighborhood of size 200. Both the clustering and the neighborhood definition

were performed using the Euclidean distance between the expression profiles.

To focus on the scoring functions and limit the impact of the optimization parameters, we first compared the number of seeds (7-mers) detected at a given FDR threshold by each scoring function. Figure 7 presents the results achieved for different number of clusters ( $k$ ), using an FDR cutoff of 0.001. For comparison, we also include the number of seeds obtained by selecting *a posteriori* the number of clusters that yields the highest number of significant seeds for each dataset (best  $k$  column). Note that this selection procedure (hereafter denoted as the best  $k$  procedure) gives a clear advantage to the original functions, since we did not allow such *a posteriori* selection of the parameters for our new scoring functions. Despite that, we observe that the new functions detect more seeds (at the same FDR) than the original ones, especially for the mutual information criterion. This shows the higher sensitivity of our approach. We also observe that the number of seeds identified by the original scoring functions strongly depends on the number of clusters  $k$ , with fewer significant seeds for the highest  $k$  values.

#### *Quality of predicted motifs*

We considered the motifs we obtained from the seeds after applying the optimization and filtering procedure described above. To evaluate the quality of these motifs, we systematically compared them to known motifs from three curated databases : JASPAR (177 motifs) [25], the collection of Gordân *et al.* (189 motifs) [26] and ScerTF (196 motifs) [27]. To determine if a predicted motif matches a known motif, we used an approach similar to that of Tomtom algorithm [28]. Briefly, given two aligned motifs, a similarity score is computed by summing (over each aligned position) the Pearson correlation coefficient between the weights of the corresponding columns in the PWMs of the corresponding motifs. For our predicted motifs, 'A' is converted to [1,0,0,0], 'W' to [0.5,0.5,0,0], etc. Using a shuffling procedure that exchanges the PWM columns of the curated database, we then estimate the p-value of obtaining an alignment of equal or higher score by chance. The best match of a predicted motif is the motif from the database which gives the highest scoring alignment, and hence the lowest p-value.

To allow a fair comparison between different motif discovery methods, we cannot strictly consider the number of predicted motifs that have a match under a given p-value, since a method which simply enumerates all the motifs would then achieve the best possible results. Hence, for each method, we took the number of predicted motifs into account and determined how many matches are significant at a *global* FDR threshold of 15%, using the procedure of Benjamini and Hochberg [29]. This way, each prediction has

an underlying cost, and the exhaustive enumeration method or a random predictor would get no significant match.

Another important issue that must be considered is the redundancy among the predictions, since a method which predicts (correctly) multiple variants of the same motif would get an unfair advantage over the other ones. Hence, when the best matches of two or more motifs correspond to the same motif in the database, only one match is considered (*i.e.* the other motif is considered as having no match).

In Figure 8 and Figure 9, we present the results obtained with ScerTF, which is the most recent and most complete database of *S. cerevisiae* motifs to date. Results obtained with the two other databases support the same conclusions and are available in Additional file 1. The left panel of these figures presents the average results for the 24 yeast datasets. Each column has two bars: the average number of predicted motifs and the average number of matches at 15% FDR. As we did for the seeds, we also included the results obtained by selecting *a posteriori* the number of clusters  $k$  that gives the highest number of motifs for each dataset (best  $k$  procedure). On average, the continuous versions of the scoring functions substantially predict more motifs than the discrete versions for any number  $k$  of clusters. Most importantly, the continuous versions also have more matches in ScerTF than the discrete versions, at 15% FDR.

To determine the significance of the observed gain, we performed a sign test between the number of matches obtained by RED<sup>2</sup> and the best  $k$  procedure for the 24 datasets (see the right panel of Figure 8 and Figure 9). The p-values of these tests are 0.0004 for the mutual information and 0.003 for the hypergeometric distribution, in favor of RED<sup>2</sup>.

When comparing Figure 7 with Figure 8 and Figure 9, we observe that more detected seeds does not necessarily lead to more motifs and more significant matches. For example, the highest average number of seeds were obtained with  $k = 5$  clusters for both scoring functions, while the highest average number of predicted motifs and matches at 15% FDR were obtained with  $k = 10$  or  $k = 20$ . This indicates that the number of detected seeds alone cannot be used as a reliable measure of performance for comparing different motif discovery approaches.

#### *Comparison to FIRE and MatrixREDUCE*

We performed a similar analysis to compare RED<sup>2</sup> (with the mutual information scoring function) to the well-established FIRE and MatrixREDUCE softwares on the same 24 datasets. Both methods were run with default parameters. For FIRE, this corresponds to a seed length of 7 bp and a maximal motif length of 9 bp, as for RED<sup>2</sup>, and we used the best  $k$  procedure described above. MatrixReduce does not allow seed

elongation as in RED<sup>2</sup> or FIRE, so we performed an optimization of all seeds of length 7, 8 and 9 bp. In both cases, we performed a double strand analysis (*i.e.* without strand distinction). Results are presented in Figure 10. First, we observe (left panel) that RED<sup>2</sup> predicts substantially more motifs, on average, than the two other methods. Second, we observe that the number of predictions that match a known motif in ScerTF at 15% FDR is significantly higher for RED<sup>2</sup> (7.75) than for FIRE (5.21) and MatrixREDUCE (3.75). This represents a 49% increase over FIRE, and a 206% increase over MatrixREDUCE. A sign test performed on the number of matches obtained for each of the 24 datasets (right panels) gives p-values of 0.0003 and 4.77e-07, for RED<sup>2</sup> against FIRE and MatrixREDUCE, respectively.

Note that the results of the above comparison should be interpreted with caution, since the three methods use different procedures to optimize the motifs, filter the output and even control the number of seeds that are optimized. For example, FIRE uses conditional mutual information to optimize each motif with regard to the already optimized motifs, while RED<sup>2</sup> independently optimizes each motif. For MatrixReduce, the differences are even greater since it is based on a linear regression model that takes the number of motif occurrences in each sequence into account, while RED<sup>2</sup> and FIRE consider only the absence or presence of a motif in each sequence. Thus, it is not possible to state objectively in which proportion the observed differences between the output of these softwares are due to the scoring functions, the clustering step (for FIRE) or the optimization procedures and parameters. Nevertheless, this comparison may be of interest to the users who are primarily concerned with the number and quality of the predicted motifs.

The motifs predicted by RED<sup>2</sup> on the Gasch *et al.* (complete compendium) and the Spellman *et al.* datasets with both scoring functions are provided in Additional files 2-5. For comparison, we also provide the list of motifs predicted by FIRE and MatrixReduce on the Spellman *et al.* dataset in Additional files 6 and 7. For each motif, these files give the most significant match in the ScerTF database (if any at 15% FDR) and the most over-represented GO term for the genes that possess the motif.

### *Orphan motifs*

All considered methods predicted some motifs that do not match any known motif according to our statistical criteria. These absences should be interpreted with caution, because determining whether a particular motif “matches” another motif is prone to error [28], and often relies on arbitrary threshold or criteria. Moreover, “high” similarity between two motifs doesn’t always translate into a significant match (and vice-versa), especially when the motifs are short and the query database contains many of them. In many cases, careful inspection suggests that these orphan motifs are sub-optimal variants of known motifs.

For example, this is clearly the case with motif HVGGDGGCR, obtained with the mutual information scoring function on the Spellman *et al.* dataset (motif #7 in Additional file 4). This motif roughly corresponds to a 3 bp shift of the motif GKGCCRAMW (motif #8) which significantly matches the RPN4 motif. In other cases, additional experiments are needed to determine if they are new biological motifs or simply false positives. For example, motif CCCTTTWHH (motif #6) does not correspond to any motif from the three considered databases. However, other information suggest that this motif may not be a false positive. First, it is strongly conserved in *S. bayanus* orthologous genes [30]. Second, the set of genes containing this motif is enriched with the “lipid modification” annotation in the Gene Ontology ( $P < 1.49e-02$ , Bonferonni corrected).

### *Complementarity of the methods*

It is worth noting that the different methods usually predicted complementary sets of motifs. For example, on the Spellman *et al.* dataset, the mutual-information and the hypergeometric scoring functions each recovered 9 motifs from the ScerTF database at 15% FDR (Additional files 4 and 5). However, only four motifs are recovered by both methods (MBP1, TOD6, SFP1, and RPN4). The hypergeometric scoring functions uniquely finds GIS1, GCN4, PHO4, REI1 and STE12, while the mutual information uniquely finds YDR026C, RAP1, MSN2, SPT15 and NDT80 (note however that GIS1 and MSN2 motifs are very similar). On the same dataset, FIRE and MatrixREDUCE find 5 and 8 motifs at 15% FDR, respectively (Additional files 6 and 7). Among them, one (GIS1) and 4 (STB1, FKH1, RTG3 and STE12) are not recovered by RED<sup>2</sup> (mutual information), respectively. Importantly, the complementarity of the methods is not restricted to individual datasets. Table 1 reports, for each pair of methods ( $m_1, m_2$ ), the number of motifs that are recovered in at least one dataset by  $m_1$ , but never by  $m_2$ . We observe, for every pair of methods, that the number of uniquely identified motifs represents an important fraction of the total number of recovered motifs. As expected, RED<sup>2</sup> with the mutual-information criterion uniquely identifies the highest number of motifs. However, FIRE also identifies several motifs that are not recovered by the other approaches (among 39 recovered motifs, 15 are not recovered by RED<sup>2</sup> with mutual-information, and 27 are not recovered by MatrixREDUCE). Moreover, although MatrixREDUCE recovers substantially less motifs (20) than the other approaches, the recovered motifs are often unique to this approach. A careful inspection suggests that these motifs often belong to a small set of genes that do not have correlated expression profiles but are over- or under-expressed in one or a few particular experiments. We provide in Additional file 8 the number of datasets in which each ScerTF motif was recovered at 15% FDR (only the

motifs recovered by at least one method are shown).

In the experiments above, we considered the number of predictions that match a known motif for a given global FDR. Since this approach takes the number of prediction into account, it is essential to have a nearly complete collection of known motifs. Otherwise, the FDR would be overestimated and many good predictions would have no significant match. In consequence, we restricted our benchmark analysis to yeast, which is the only eukaryote for which such collections exist. Nevertheless, we also applied RED<sup>2</sup> to several Apicomplexan parasites (see the application to *P. falciparum* below), as well as to human and *Arabidopsis Thaliana*. For this latter, RED<sup>2</sup> has been successfully used to detect DNA motifs involved in transcriptional regulation following phosphate starvation and specific drug treatment [31].

### **Application to *P. falciparum***

*P. falciparum*, the main causal agent of human malaria, kills nearly 800000 people yearly in the 106 malaria-endemic countries [32]. Its genome is very atypical, with around 60% of the  $\sim 5\,500$  predicted genes with no sufficient similarity to characterized genes in other species to justify provision of functional assignments [33]. The way the parasite controls the expression of its genes is also poorly understood. While several microarray studies have revealed intricate and tight gene expression regulation, we are unaware of the mechanisms underlying this control, and, more specifically, of the relative role of transcriptional and post-transcriptional regulation in the parasite [34]. Moreover, with the notable exception of the recently discovered ApiAP2 domain [35], most attempts to identify transcriptional factors in *P. falciparum* have failed.

As a case study, we ran RED<sup>2</sup> on the datasets of Bozdech et al. [11] and Shock et al. [36]. The former measures mRNA levels of *P. falciparum* genes once per hour during one complete 48 h intraerythrocytic cycle, while the latter measures mRNA-decay profiles during this same cycle. For each dataset, we performed two separate analyses. In the first one, we searched for putative motifs involved in mRNA synthesis regulation, *i.e.* located up to 1 kb upstream of the start codon, either on the forward or the reverse strand. In the second one, we searched for motifs that are likely involved in mRNA degradation or sequestration, *i.e.* located up to 500 bp downstream of the stop codon, on the forward strand only, which approximately corresponds to the 3'UTR. RED<sup>2</sup> was run with the mutual information scoring function, and the following parameters :  $\alpha = 0.9$ ,  $\gamma = 0.75$ , FDR threshold = 0.001. For the downstream analysis, we also increased the maximum motif length to 12. Complete results for these analyses are provided in

Additional files 9 and 10.

The upstream analysis of the Bozdech gene expression dataset yields a total of 21 motifs. Eight of these motifs match (at 15% FDR) one of the 23 ApiAP2 motifs previously identified in a protein binding microarray experiment [18]. Among the motifs that do not match any ApiAP2 motifs, the HAGACA motif (see Figure 4(b)) achieves a very high score (it is the second best scoring motif of the analysis) and may be the binding site of a yet uncharacterized transcription factor. Downstream analysis on the same dataset identifies no motif. For information, FIRE and MatrixREDUCE (default parameters) get 6 and 3 matches (at 15% FDR) on this dataset, respectively (see Additional files 11 and 12).

In contrast, downstream analysis of the Shock mRNA decay dataset returns 6 motifs. The best scoring one (HWKTTTTTNGT, see Figure 11(a)) is partially included in a 47 bp motif previously identified in the 3'UTR of nine translationally repressed transcripts [37], and is very similar to a motif (TYTTTTNGT) identified in the downstream region of some *P. falciparum* genes by comparative genomics with *P. yoelii* and *P. knowlesi* [30]. To the best of our knowledge, the second best scoring motif (AAAAAAAAAAAV, Figure 11(b)) does not resemble any other motif already described in the literature. When looking at the occurrence position distributions, we see that both motifs have a high, although very different, positional bias. We can also see that the genes where these motifs are present have quite opposite mRNA decay profiles, and that these two motifs seem to define two different gene classes. The three remaining motifs achieve lower scores, and look similar to the best two motifs, being either T- or A-rich. Moreover, a visual inspection of their position distribution in the sequence and in the expression space shows that they are also similarly distributed. Hence, it is tempting to speculate that these motifs are variants of the best two motifs, or are complementary motifs working together with these latter.

Interestingly, a downstream analysis with the Bozdech dataset does not return any motif. Hence, the fact that these motifs were identified only in the downstream analysis of the Shock dataset supports the hypothesis that they are involved in mRNA degradation or sequestration. Moreover, it is worth noting that the downstream analysis enable the discovery of few motifs (5) compared with the equivalent upstream analysis (> 20). Together, these results seem to indicate that either transcript regulation in *P. falciparum* mainly occurs at the mRNA synthesis level, or that important downstream motifs cannot be represented as consensus sequences (e.g. structural motifs).

The RED<sup>2</sup> analysis results for yeast and *P. falciparum* have several differences. In *S. cerevisiae*, many motifs show a strong positional bias towards a small region located approximately 150 bp upstream of the

genes (e.g. SFP1, TOD6, MBP1, YDR026C and STB1, see Additional files 2-5). Interestingly, this kind of positional bias is not observed in the upstream analysis of *P. falciparum*, where motif occurrences are more uniformly distributed (see for example Additional file 9). Moreover, there is a strong strand bias for many motifs found in *P. falciparum*—the p-value of the sign test is below  $10^{-3}$  for 6 motifs—but very few in *S. cerevisiae*. Finally, we note that the average number of motif occurrences per gene is significantly higher in *P. falciparum* than in *S. cerevisiae*. Overall, these results may reflect marked differences between the gene regulation mechanisms in the two species.

## Conclusions

Most approaches for *in silico* discovery of regulatory elements require clustering of the expression data. This implicitly assumes that genes belong to disjoint classes, an assumption that rarely holds in the regulation context. In this paper, we have presented a new approach for discovering regulatory elements that does not require clustering and allows overlaps between different gene sets. This is done by estimating motif densities locally around each gene using a fixed-size neighborhood. We have adapted the hypergeometric and mutual information criteria to this approach, and have shown on yeast that the new versions outperform the original methods, both in terms of statistical sensitivity and the number of experimentally validated motifs they retrieve. A comparison with the FIRE and MatrixREDUCE algorithms also demonstrated the complementarity of the different approaches.

Furthermore, we have shown that motif densities in the gene-expression space are useful for comparing and discriminating between different motifs. This discriminative feature is extensively used in our optimization procedure to avoid merging motifs whose descriptions look similar at the sequence level, but which are associated with different locations in the expression space. Similarly, the density of a motif around a particular gene can help to predict if an occurrence of this motif is biologically active or not.

The scoring functions and methods presented in this paper have been implemented in the RED<sup>2</sup> software, which can be run from <http://www.atgc-montpellier.fr/RED2/>. RED<sup>2</sup> returns the identified motifs along with other information and figures in a user-friendly interface. A RED<sup>2</sup> analysis usually takes between 5 and 20 min., depending on the number of screened genes and the size of the analysed sequences. The source code of the program is available at <https://bitbucket.org/mlajoie/red2>.

There are several prospects for this work. First, we intend to enrich the RED<sup>2</sup> server with additional pre-loaded species and tools. For example, we plan to add a motif comparison tool that would make use of

both DNA sequence and gene expression to compute a similarity score between two motifs, in a way similar to that of our filtering procedure. Finally, another interesting prospect would be to supplement RED<sup>2</sup> with information on close species in order to improve both its sensitivity and accuracy.

## Material and methods

### Neighborhood definition

Let  $\mathcal{G}$  be the  $k$ -nearest neighbor graph obtained by putting an arc from the node (gene)  $g$  to the node  $g'$  if  $g'$  is among the  $k$ -nearest neighbors of  $g$ . In this graph, the neighborhood of a gene  $g$  involves all genes  $g'$  such that there is an arc  $g \rightarrow g'$ . Firstly, the neighborhood relations are symmetrized: if there is an arc  $g \rightarrow g'$  and no arc  $g' \rightarrow g$ , this latter is added to the graph. This gives us a new graph  $\mathcal{G}^*$ . In this graph, each gene is in the neighborhood of at least  $k$  other genes. We use  $\mathcal{G}^*$  to progressively construct  $\mathcal{G}_k$ , a directed graph where each gene has exactly  $k$  neighbors, in a way that reduces the skewness of the in-degree distribution. More precisely, the  $k$  neighbors of each gene  $g$  are successively sampled from its neighbors in  $\mathcal{G}^*$ , according to the following distribution

$$P(g') = \left[ \sum_{g'' \in \mathcal{N}(g)} \frac{d(g')}{d(g'')} \right]^{-1},$$

where  $d(g')$  is the in-degree of gene  $g'$  in  $\mathcal{G}^*$ , and  $\mathcal{N}(g)$  the set of neighbors of  $g$  in  $\mathcal{G}^*$  that have not yet been included in  $\mathcal{G}_k$ . In other words, the probability for a gene to be chosen as a neighbor is inversely proportional to the number of times it can be chosen.

### Motif density

The density of a motif in a subset of genes is simply the proportion of genes that contain the motif in that subset. To better distinguish between high and low density values, it is sometimes convenient to express the density as a z-score. Let  $n_m$  be the number of genes that contain a given motif  $m$  among a total of  $n$  genes. Under  $H_0$  (random motif distribution), the number of genes that contain the motif among a sample of  $k$  genes ( $n_{mk}$ ) follows a hypergeometric distribution [38] of parameters  $n$ ,  $n_m$  and  $k$ . Thus, the z-score of an observed density is given by

$$\frac{kn_m/n}{\sqrt{\frac{kn_m(n-k)(n-n_m)}{n^2(n-1)}}} \quad (6)$$

This quantity is positive (resp. negative) if the number of genes that possess the motif in the sample is higher (resp. lower) than expected.

## Yeast and *P. falciparum* data

The 1 kb and 600 bp upstream sequences of *P. falciparum* and *S. cerevisiae* (respectively) were downloaded from the FIRE website [6]. The 500 bp downstream sequences of *P. falciparum* were downloaded from the PlasmoDB database [39] (version 6.3). To prevent any bias that could result from multi-gene families [6], paralogs were identified using OrthoMCL [40] and discarded.

The Gasch et al. compendium was divided into 22 subsets according to the descriptions available in the file. The resulting partition is available in Additional file 13. Each data then consists in expression profiles of size between 3 to 15 time-points/experiments, with an average number of 7.5 time-points/experiments per data.

## Competing interests

The authors declare that they have no competing interests.

## Authors contributions

ML and LB designed the method and experiments, and drafted the manuscript. ML conceived and implemented the program and conducted the experiments. OG suggested method improvements and revised the manuscript. VL integrated the program into the web server and make technical improvements. LB coordinated the study. All authors read and approved the final manuscript.

## Acknowledgements

This research is supported by the French National Research Agency: PlasmoExplore project (ANR-06-CIS6-MDCA-014) and PlasmoExpress project (ANR JCJC 2010).

## References

1. Mello C, Conte D: **Revealing the world of RNA interference.** *Nature* 2004, **431**:338–342.
2. Bailey T, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** In *ISMB. International Conference on Intelligent Systems for Molecular Biology, Volume 2* 1994:28.
3. Hughes J, Estep P, Tavazoie S, Church G: **Computational identification of Cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*.** *Journal of Molecular Biology* 2000, **296**:1205–1214.
4. Pavesi G, Mauri G, Pesole G: **An algorithm for finding signals of unknown length in DNA sequences.** *Bioinformatics* 2001, **17**:S207.
5. Sandve G, Abul O, Walseng V, Drabløs F: **Improved benchmarks for computational motif discovery.** *BMC bioinformatics* 2007, **8**:193.

6. Elemento O, Slonim N, Tavazoie S: **A universal framework for regulatory element discovery across all genomes and data types.** *Molecular cell* 2007, **28**:337–350.
7. Young J, Johnson J, Benner C, Yan S, Chen K, Le Roch K, Zhou Y, Winzeler E: **In silico discovery of transcription regulatory elements in Plasmodium falciparum.** *BMC Genomics* 2008, **9**:70.
8. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan MS, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin, et al.: **Genome sequence of the human malaria parasite Plasmodium falciparum.** *Nature* 2002, **419**:498–511.
9. Wagner A: **Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes.** *Bioinformatics* 1999, **15**:776.
10. Hobert O: **Gene regulation by transcription factors and microRNAs.** *Science* 2008, **319**:1785.
11. Bozdech Z, Llinás M, Pulliam B, Wong E, Zhu J, DeRisi J: **The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum.** *PLoS Biol* 2003, **1**:E5.
12. Chen X, Hughes T, Morris Q: **RankMotif++: a motif-search algorithm that accounts for relative ranks of K-mers in binding transcription factors.** *Bioinformatics* 2007, **23**:i72.
13. Eden E, Lipson D, Yogev S, Yakhini Z: **Discovering motifs in ranked lists of DNA sequences.** *PLoS Comput Biol* 2007, **3**:e39.
14. Bussemaker H, Li H, Siggia E: **Regulatory element detection using correlation with expression.** *Nature genetics* 2001, **27**:167–174.
15. Foat B, Houshmandi S, Olivas W, Bussemaker H: **Profiling condition-specific, genome-wide regulation of mRNA stability in yeast.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**:17675.
16. Holmes I, Bruno W: **Finding regulatory elements using joint likelihoods for sequence and expression profile data.** In *Proc. Int. Conf. Intell. Syst. Mol. Biol, Volume 8* 2000:202–210.
17. De Silva E, Gehrke A, Olszewski K, León I, Chahal J, Bulyk M, Llinás M: **Specific DNA-binding by apicomplexan AP2 transcription factors.** *Proceedings of the National Academy of Sciences* 2008, **105**:8393.
18. Campbell T, De Silva E, Olszewski K, Elemento O, Llinás M, Smith J: **Identification and Genome-Wide Prediction of DNA Binding Specificities for the ApiAP2 Family of Regulators from the Malaria Parasite.** *PLoS Pathogens* 2010, **6**:5–526.
19. Riordan D, Herschlag D, Brown P: **Identification of RNA recognition elements in the Saccharomyces cerevisiae transcriptome.** *Nucleic Acids Research* 2010.
20. Radovanović M, Nanopoulos A, Ivanović M: **Hubs in space: Popular nearest neighbors in high-dimensional data.** *The Journal of Machine Learning Research* 2010, **11**:2487–2531.
21. Gasch A, Spellman P, Kao C, Carmel-Harel O, Eisen M, Storz G, Botstein D, Brown P: **Genomic expression programs in the response of yeast cells to environmental changes.** *Science's STKE* 2000, **11**:4241.
22. Nguyen D, D'haeseleer P: **Deciphering principles of transcription regulation in eukaryotic genomes.** *Molecular systems biology* 2006, **2**.
23. Spellman P, Sherlock G, Zhang M, Iyer V, Anders K, Eisen M, Brown P, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.** *Molecular biology of the cell* 1998, **9**:3273.
24. R Development Core Team: *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria 2011, [http://www.R-project.org/]. [ISBN 3-900051-07-0].
25. Portales-Casamar E, Thongjuea S, Kwon A, Arenillas D, Zhao X, Valen E, Yusuf D, Lenhard B, Wasserman W, Sandelin A: **JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles.** *Nucleic acids research* 2010, **38**:D105.
26. Gordan R, Murphy K, McCord R, Zhu C, Vedenko A, Bulyk M: **Curated collection of yeast transcription factor DNA binding specificity data reveals novel structural and gene regulatory insights.** *Genome Biology* 2011, **12**:R125.

27. Spivak A, Stormo G: **ScerTF: a comprehensive database of benchmarked position weight matrices for *Saccharomyces* species.** *Nucleic Acids Research* 2012, **40**:D162–D168.
28. Gupta S, Stamatoyannopoulos J, Bailey T, Noble W: **Quantifying similarity between motifs.** *Genome biology* 2007, **8**:R24.
29. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *Journal of the Royal Statistical Society. Series B (Methodological)* 1995, :289–300.
30. Wu J, Sieglaff D, Gervin J, Xie X: **Discovering regulatory motifs in the *Plasmodium* genome using comparative genomics.** *Bioinformatics* 2008, **24**:1843.
31. Boudière L, Botté C, Saidani N, Lajoie M, Marion J, Bréhélin L, Yamaryo-Botté Y, Satiat-Jeunemaître B, Breton C, Girard-Egrot A, Bastien O, Jouhet J, Falconet D, Block M, Maréchal E: **Galvestine-1, a novel chemical probe for the study of the glycerolipid homeostasis system in plant cells.** *Mol. BioSyst.* 2012.
32. World Health Organization: **World Malaria Report** 2010.
33. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan MS, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DM, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian GM, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM, Barrell B: **Genome sequence of the human malaria parasite *Plasmodium falciparum*.** *Nature* 2002, **419**:498–511.
34. Horrocks P, Wong E, Russel K, Emes R: **Control of gene expression in *Plasmodium falciparum* - Ten years on.** *Molecular & Biochemical Parasitology* 2009, **164**:9–25.
35. Balaji S, Babu M, Iyer L, Aravind L: **Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains.** *Nucleic acids research* 2005, **33**:3994.
36. Shock J, Fischer K, DeRisi J: **Whole-genome analysis of mRNA decay in *Plasmodium falciparum* reveals a global lengthening of mRNA half-life during the intra-erythrocytic development cycle.** *Genome Biology* 2007, **8**:R134.
37. Hall N, Karras M, Raine JD, Carlton JM, Kooij TW, Berriman M, Florens L, Janssen CS, Pain A, Christophides GK, James K, Rutherford K, Harris B, Harris D, Churcher C, Quail MA, Ormond D, Doggett J, Trueman HE, Mendoza J, Bidwell SL, Rajandream MA, Carucci DJ, Yates JR, Kafatos FC, Janse CJ, Barrell B, Turner CM, Waters AP, Sinden RE: **A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses.** *Science* 2005, **307**:82–86".
38. Ewens W, Grant G: *Statistical methods in bioinformatics: an introduction, Volume 10.* Springer Verlag 2005.
39. Aurrecochea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, Heiges M, Innamorato F, Iodice J, Kissinger JC, Kraemer E, Li W, Miller JA, Nayak V, Pennington C, Pinney DF, Roos DS, Ross C, Stoeckert CJ, Treatman C, Wang H: **PlasmoDB: a functional genomic database for malaria parasites.** *Nucleic Acids Res.* 2009, **37**(Database issue):D539–543.
40. Li L, Stoeckert C, Roos D: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome research* 2003, **13**:2178.

## Figures

**Figure 1** — Gene expression data of *P. falciparum* 48 h erythrocytic cycle (Bozdech et al. (2003)). Each point in the figures corresponds to an expression profile plotted according to its two principal components. All points remain at the same coordinate through the six figures (*i.e.* only the coloring changes). In the top row figures, colors indicate the cluster membership of each profile after  $k$ -means clustering of the original data (*i.e.* prior to dimensionality reduction) using Pearson correlation distance for  $k = 3, 5$  and  $7$ . We observe that the clusters are nearly equally-sized and that their edges are rather arbitrary, as they do not follow low density regions and change radically for different  $k$  values. Figures in the bottom row show regions of the expression space enriched for three regulatory motifs identified in previous studies [6, 7, 18]. Enrichment is defined on the original data (*i.e.* prior to dimensionality reduction) by measuring the proportion of genes that possess the motif in their upstream sequence (1kb) among the 200 nearest neighbors of each gene (according to their profiles). Colored points correspond to profiles where this proportion is 3 standard deviations above the expected one according to the hypergeometric law (see motif density in Material and methods). Non colored points were shaded for clarity. We observe that 1) each motif corresponds to a contiguous region of the expression space, 2) these regions do not correspond to those defined by any of the above clustering and 3) regions defined by different motifs can strongly overlap, highlighting the weaknesses of clustering-based approaches for motif discovery.

**Figure 2** — In-degree distribution of two different neighborhood graphs for the Gasch et al. *S. cerevisiae* gene expression compendium [21]. (Left)  $k$  nearest-neighbors graph ( $k=200$ ). (Right) Our modified  $k$ -neighbors graph ( $k=200$ ).

**Figure 3** — Comparison between different motifs in *P. falciparum*, using the Pearson correlation between their density profiles (Bozdech et al. dataset). Each point corresponds to a gene and represents a pair of density estimates (one for each motif). The left figure suggests that the AAGACA and TAGACA sequences belong to the same motif (WAGACA), as their density profiles have a Pearson correlation coefficient of 0.94. On the contrary, the TAGACA and TACACA sequences have clearly different density profiles, despite the fact that they differ by only one nucleotide, as shown in the right figure, with a Pearson coefficient of only 0.54. The densities are estimated using 200 neighbors for each gene and are expressed as a z-score (see motif density in Material and methods).

**Figure 4** — The three highest scoring motifs found in the Bozdech analysis. (Left) Motif logos (Middle) Distance distributions according to the start codon (Right) Profile heatmaps. The x-axis of this map represents the different conditions in the expression dataset, while the y-axis represents the expression level. The y-axis is divided into 25 bins, and the colors of the map indicate the motif density (expressed as a z-score) in each of these bins (red: high density; green: low density).

**Figure 5** — Relation between motif density and functionality. Each point corresponds to a PAC/TOD6 motif occurrence (HGMGATGAG) in a promoter region of a *S. cerevisiae* gene. The x-axis represents the distance in bp of an occurrence from the start codon. The y-axis represents the motif density around the gene where the occurrence is found. Density is estimated by considering the number of genes containing the PAC motif in the neighborhood of each gene (Gasch et al. dataset), and is expressed as a z-score (see motif density in Material and methods).

**Figure 6** — Effect of the neighborhood definition on the scoring function with the Gasch et al. dataset. The Y-axis is the number of significant seeds (7-mers) detected at different *FDR* thresholds (X-axis). (Left) Mutual information scoring function. (Right) Hypergeometric scoring function.

**Figure 7** — Average number of seeds detected by the continuous and discrete versions of each scoring functions for the 24 yeast datasets ( $FDR < 0.001$ ). (Left) Mutual information. (Right) Hypergeometric. For each panel, the leftmost column shows the results obtained by the continuous version implemented in RED<sup>2</sup>. The right columns show the results obtained by the discrete version for different numbers ( $k$ ) of clusters. The “best  $k$ ” column shows the average results when the clustering that yields the highest number of detected seeds is selected *a posteriori* for each dataset.

**Figure 8** — Average number of motifs and matches at 15% FDR recovered by the continuous and discrete mutual-information scoring functions. The same optimization and filtering procedure was applied on both versions (seed  $FDR < 0.001$ ,  $\alpha = 0.75$ ,  $\gamma = 0.75$ ). (Left) Average results for the 24 yeast datasets. The leftmost column shows the results obtained by the continuous version of the scoring function implemented in RED<sup>2</sup>. The “ $k=i$ ” columns show the results obtained by the discrete version and  $i$  clusters. The “best  $k$ ” column shows the results obtained when the number of clusters that yields the highest number of motifs is selected *a posteriori* for each dataset. (Right) Number of predicted motifs that match a known motif in the ScerTF database for the 24 yeast datasets ( $FDR < 15\%$ ). Each point corresponds to one of the 24 datasets. The y-axis corresponds to the number achieved by RED<sup>2</sup> and the x-axis to the number achieved by the discrete version and the best  $k$  procedure. Superimposed points are indicated by a shading. RED<sup>2</sup> found more motifs than the best clustering for 18 datasets and less for 2 datasets, which gives a sign test p-value of 0.0004.

**Figure 9** — Average number of motifs and matches at 15% FDR recovered by the continuous and discrete hypergeometric-distribution scoring functions. The same optimization and filtering procedure was applied on both versions (seed  $FDR < 0.001$ ,  $\alpha = 0.75$ ,  $\gamma = 0.75$ ). (Left) Average results for the 24 yeast datasets. The leftmost column shows the results obtained by the continuous version of the scoring function implemented in RED<sup>2</sup>. The “ $k=i$ ” columns show the results obtained by the discrete version and  $i$  clusters. The “best  $k$ ” column shows the results obtained when the number of clusters that yields the highest number of motifs is selected *a posteriori* for each dataset. (Right) Comparison of the number of predicted motifs that match a known motif in the ScerTF database for the 24 yeast datasets at 15% FDR. The y-axis corresponds to the number achieved by RED<sup>2</sup> and the x-axis to the number achieved by the discrete version and the best  $k$ . Superimposed points are

indicated by a shading. RED<sup>2</sup> found more motifs than the best clustering for 17 datasets and less for 3 datasets, which gives a sign test p-value of 0.003.

**Figure 10** — Number of predicted motifs and number of matches in ScerTF at 15% FDR, for RED<sup>2</sup>(mutual-information scoring function; same as Figure 8), FIRE and MatrixREDUCE. FIRE was run with default parameters (optimized for yeast), with  $k = 5, 10, 20, 40$  and 80 clusters, and the number of clusters that yields the highest number of motifs was selected *a posteriori* for each dataset. MatrixREDUCE was run with default parameters, with seeds of length 7, 8 and 9. (Left) Average results of the three methods on the 24 yeast datasets. (Center) RED<sup>2</sup>and FIRE. Number of predicted motifs that match a known motif at 15% FDR in the ScerTF database for the 24 yeast datasets. The y-axis corresponds to the number achieved by RED<sup>2</sup> and the x-axis to the number achieved by FIRE with the best clustering procedure. Superimposed points are indicated by a shading. RED<sup>2</sup> has more matches than FIRE in 21 datasets and less in 3 datasets, which gives a sign test p-value of 0.0003. (Right) RED<sup>2</sup>and MatrixREDUCE (same explanations as for the center panel). RED<sup>2</sup> has more matches than MatrixREDUCE for 22 datasets and less for no dataset, which gives a sign test p-value of 4.77e-07

**Figure 11** — The two highest scoring motifs found in the Shock analysis, using the mutual information scoring function. (Left) Motif logos (Middle) Distance distributions according to the stop codon (Right) Profile heatmaps.

## Tables

**Table 1 - Number of motifs recovered and not recovered by the different methods at 15% FDR**

Each cell  $(i, j)$  provides the number of ScerTF motifs recovered (in at least one of the 24 datasets) by method  $i$ , but not recovered by method  $j$ . The total number of motifs recovered by each method is given in parentheses (first column). For example, among the 37 motifs recovered by RED<sup>2</sup> and the hypergeometric scoring function, 13 are not recovered by FIRE.

	RED2-MI	RED2-H	FIRE	REDUCE
RED2-MI (45)	-	21	21	35
RED2-H (37)	13	-	13	25
FIRE (39)	15	15	-	28
REDUCE (20)	10	8	9	-

## Additional files

### Additional file 1

Title: Assessment on *S. cerevisiae* using the JASPAR and the Gordán *et al.* databases.

Description: Results obtained when using the JASPAR and the Gordán *et al.* databases (instead of ScerTF). This corresponds to the experiments shown in Figures 8-10 in the main manuscript.

### **Additional file 2**

Title: Results of RED<sup>2</sup> with mutual information criterion on the *S. cerevisiae* upstream regions with the Gasch et al. stress compendium.

Description: The set of motifs inferred by RED<sup>2</sup> on the Gasch et al. complete compendium. Provided for each motif are the logo of the motif, the number of genes that have the motif in their promoter region, the expression profiles (heatmap) associated with the motif, an histogram of the motif occurrence positions, the existence/absence of a strand bias, and the name of the ScerTF motif that best matches the inferred motif (if any at 15% FDR).

### **Additional file 3**

Title: Results of RED<sup>2</sup> (hypergeometric) on *S. cerevisiae* upstream regions with the Gasch et al. stress compendium.

Description: The set of motifs inferred by RED<sup>2</sup> on the Gasch et al. complete compendium. See the description of Additional file 2 for a description of the different columns.

### **Additional file 4**

Title: Results of RED<sup>2</sup> (mutual information) on *S. cerevisiae* upstream regions with the Spellman et al. cell-cycle dataset.

Description: The set of motifs inferred by RED<sup>2</sup> on the Spellman *et al.* dataset. See the description of Additional file 2 for a description of the different columns.

### **Additional file 5**

Title: Results of RED<sup>2</sup> (hypergeometric) on *S. cerevisiae* upstream regions with the Spellman et al. cell-cycle dataset.

Description: The set of motifs inferred by RED<sup>2</sup> on the Spellman *et al.* dataset. See the description of Additional file 2 for a description of the different columns.

### **Additional file 6**

Title: Results of FIRE on *S. cerevisiae* upstream regions with the Spellman et al. cell-cycle dataset.

Description: The set of motifs inferred by FIRE on the Spellman *et al.* dataset. See the description of Additional file 2 for a description of the different columns.

### **Additional file 7**

Title: Results of MatrixREDUCE on *S. cerevisiae* upstream regions with the Spellman et al. cell-cycle dataset.

Description: The set of motifs inferred by MatrixREDUCE on the Spellman *et al.* dataset. See the description of Additional file 2 for a description of the different columns.

### **Additional file 8**

Title: ScerTF motifs recovered by RED<sup>2</sup>, FIRE and MatrixREDUCE on the 24 *S. cerevisiae* datasets.

Description: For each method, this table provides the number of datasets in which each motif was recovered (at 15% FDR). For example, REDUCE correctly predicts AFT1 (first line) in one dataset. The last two lines give the total number of correct predictions (Total) and the number of distinct motifs recovered (Coverage). For example, on the 24 datasets, RED2 makes 177 correct predictions, which covers 43 distinct motifs in ScerTF.

### **Additional file 9**

Title: Results of RED<sup>2</sup> (mutual information) on *P. falciparum* upstream regions with the Bozdech et al. dataset (erythrocytic cycle).

Description: The set of motifs inferred by RED<sup>2</sup> on the upstream regions of *P. falciparum* genes using the Bozdech et al. dataset. See the description of Additional file 2 for a description of the different columns.

### **Additional file 10**

Title: Results of RED<sup>2</sup> (mutual information) on *P. falciparum* downstream regions with the Shock et al. dataset (mRNA decay).

Description: The set of motifs inferred by RED<sup>2</sup> on the downstream regions of *P. falciparum* genes using the Shock et al. dataset. See the description of Additional file 2 for a description of the different columns.

### **Additional file 11**

Title: Results of FIRE on *P. falciparum* upstream regions with the Bozdech et al. dataset (erythrocytic cycle).

Description: The set of motifs inferred by FIRE on the upstream regions of *P. falciparum* genes using the Bozdech et al. dataset. See the description of Additional file 2 for a description of the different columns.

**Additional file 12**

Title: Results of MatrixREDUCE on *P. falciparum* upstream regions with the Bozdech et al. dataset (erythrocytic cycle).

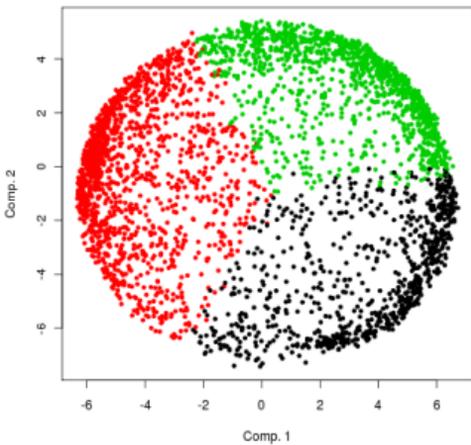
Description: The set of motifs inferred by MatrixREDUCE on the upstream regions of *P. falciparum* genes using the Bozdech et al. dataset. See the description of Additional file 2 for a description of the different columns.

**Additional file 13**

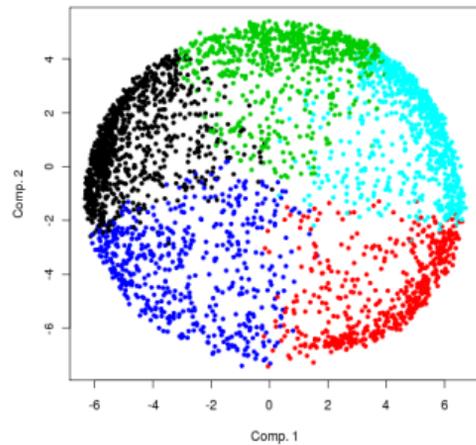
Title: The 22 datasets from the Gasch et al. compendium.

Description: This text file describes the partition of the Gasch et al. stress compendium into 22 test datasets.

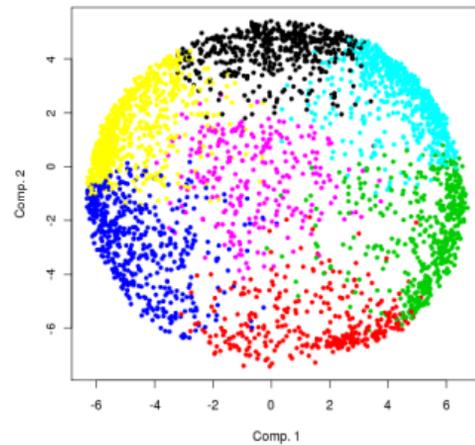
**K-means, k=3**



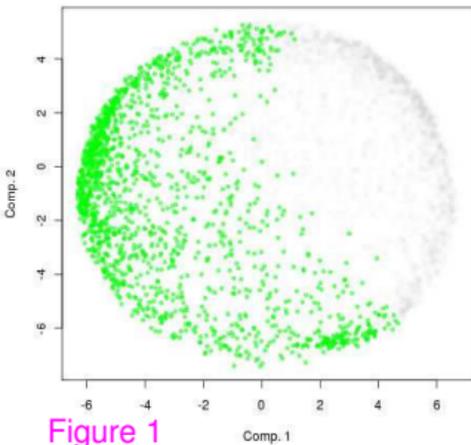
**K-means, k=5**



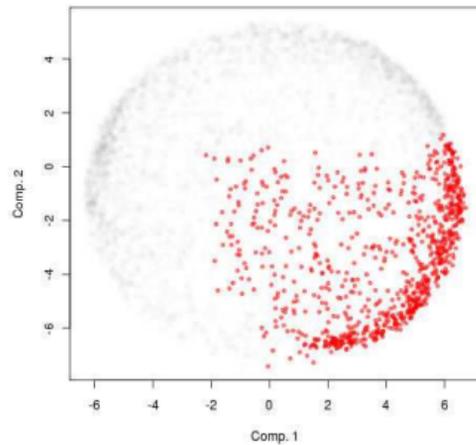
**K-means, k=7**



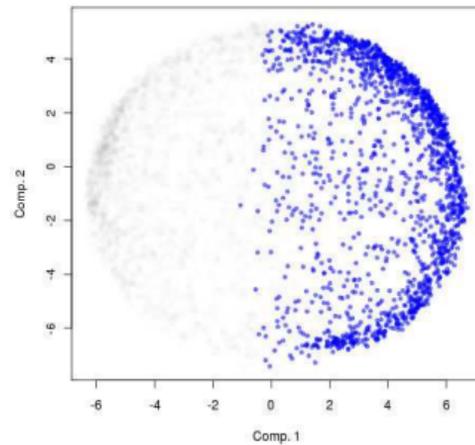
**DHTRTAGAD**



**DWAGACA**

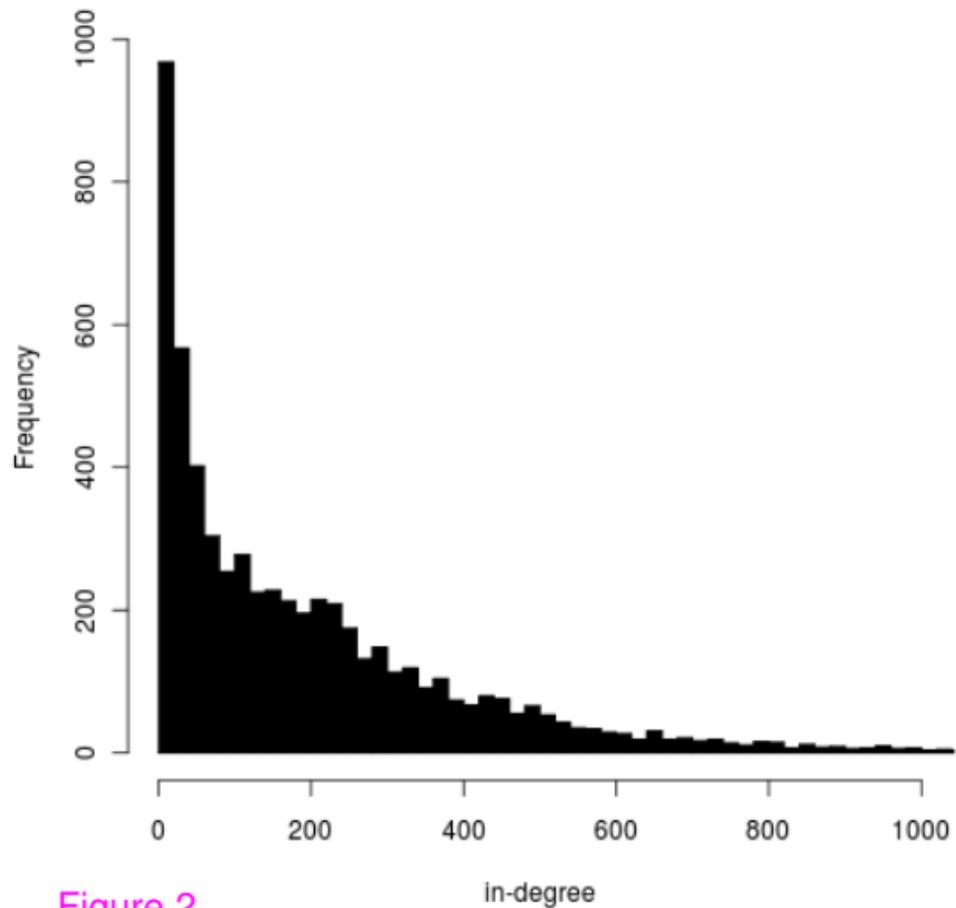


**DRTGTGYR**



**Figure 1**

kNN graph



modified k-neighbors graph

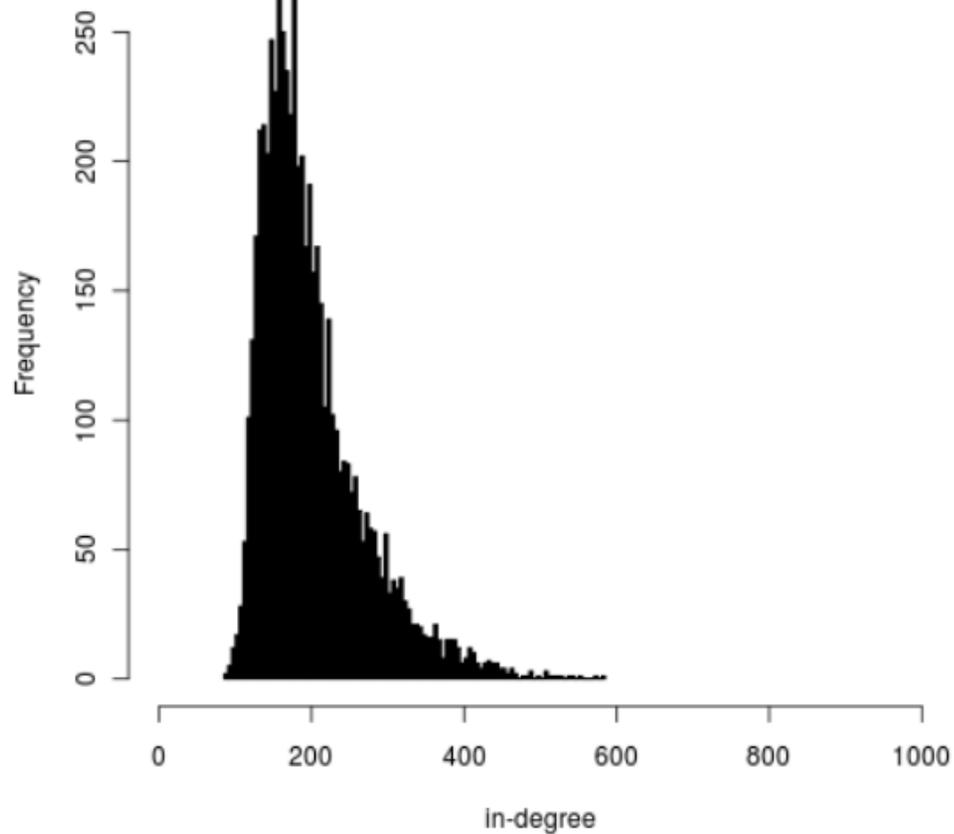
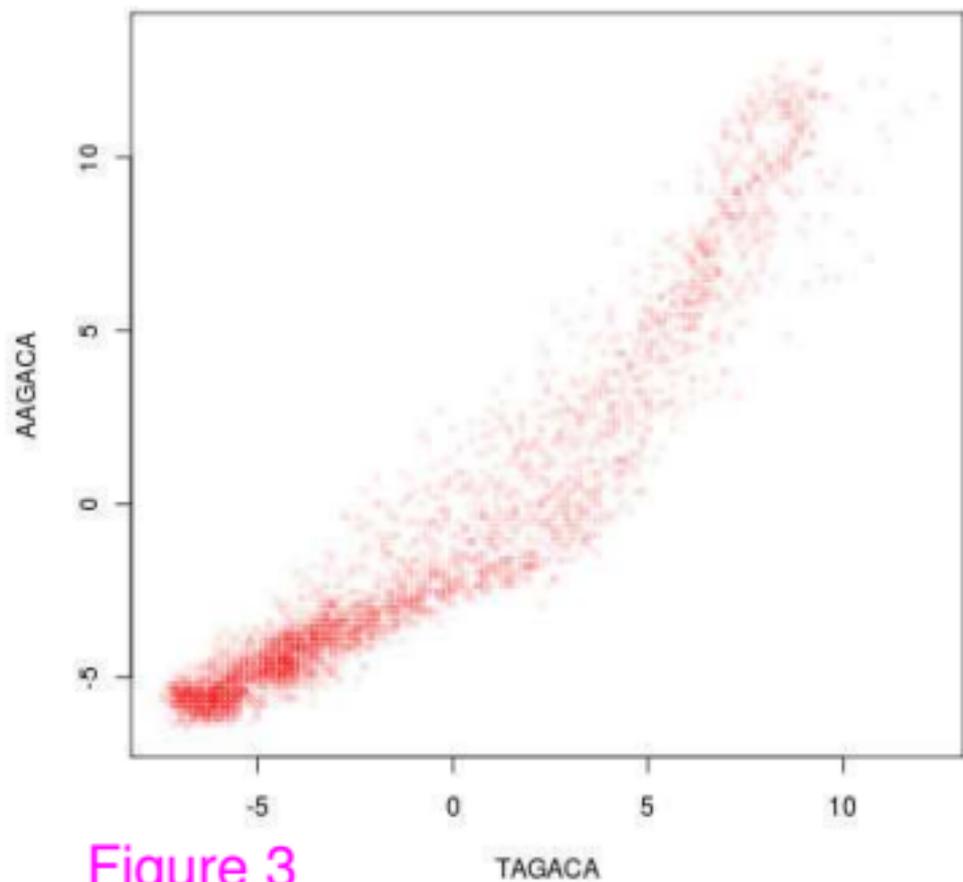


Figure 2

**AAGACA vs TAGACA**  
 $r = 0.94$



**TAGACA vs TACACA**  
 $r = 0.54$

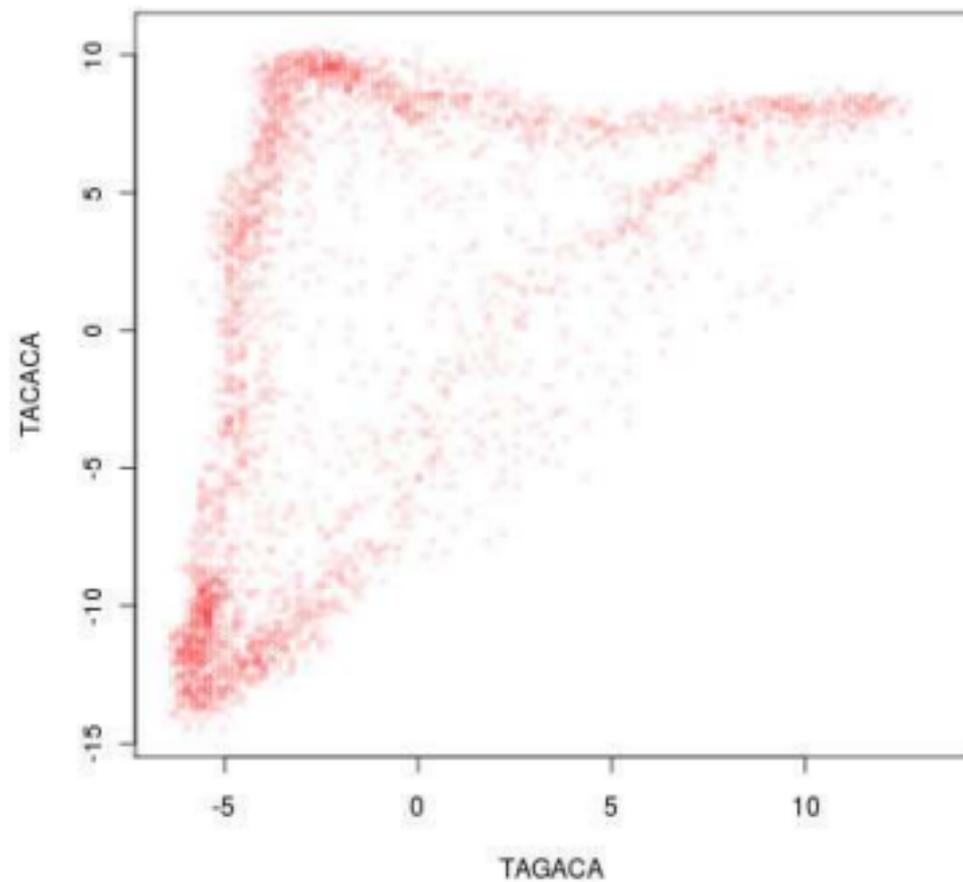
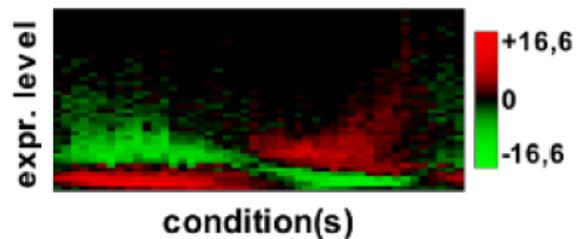
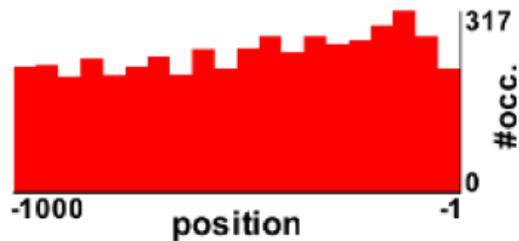


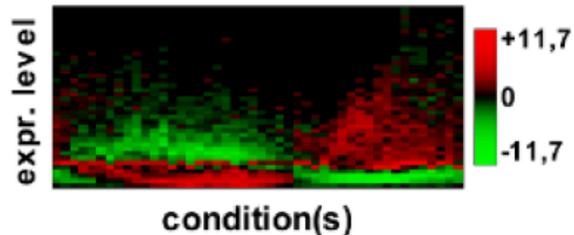
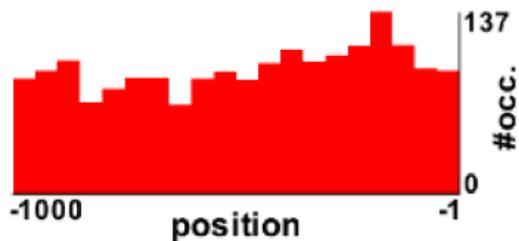
Figure 3

CA CACA C  
TG T



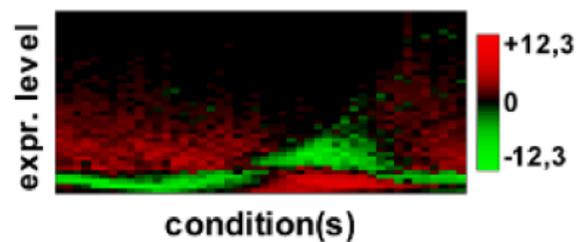
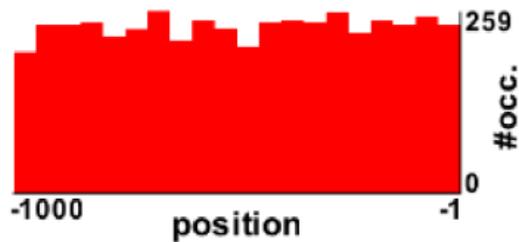
(a)

AGACA



(b)

T TAGA  
T G T



(c)

Figure 4

## PAC motif position vs density

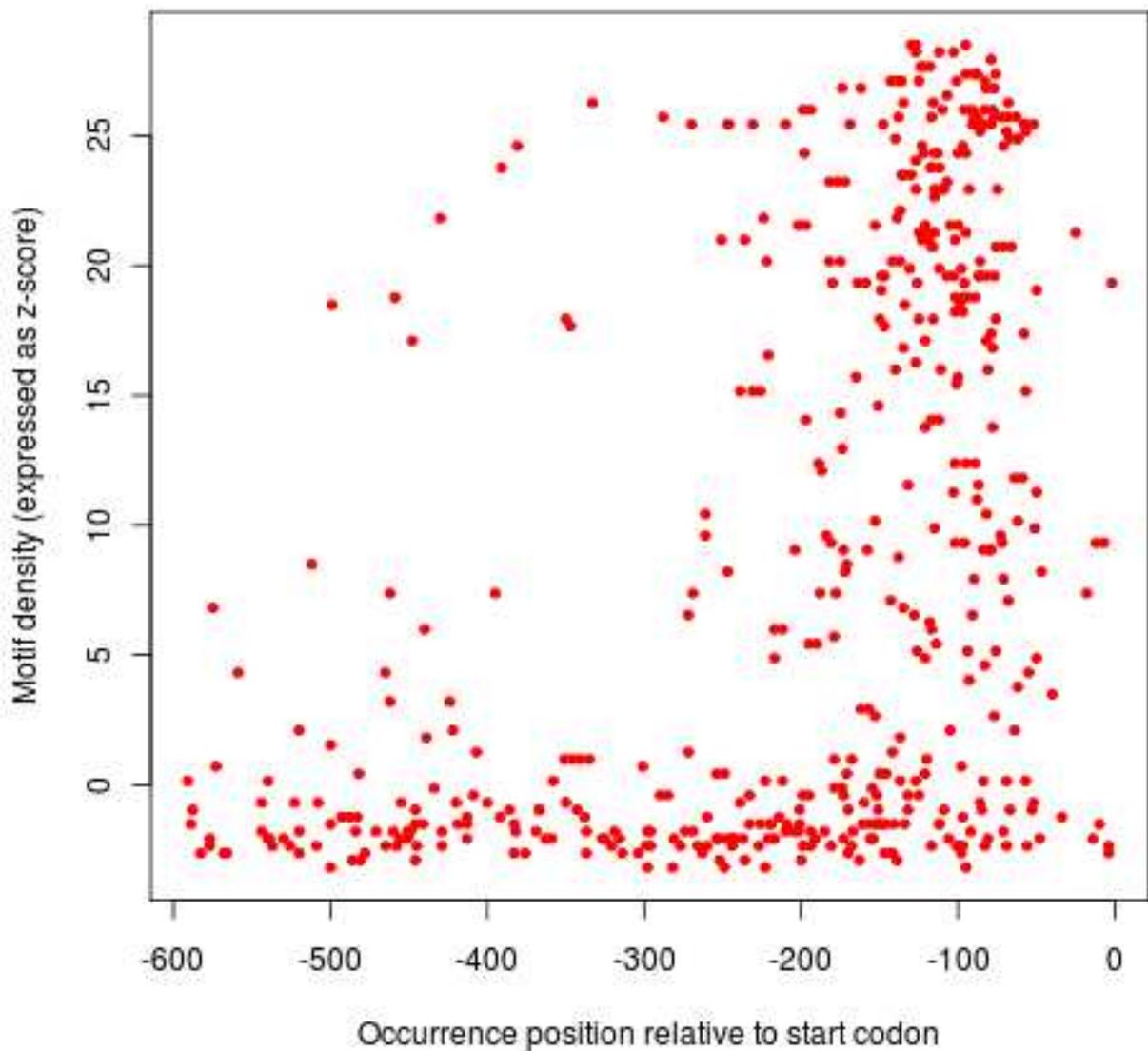
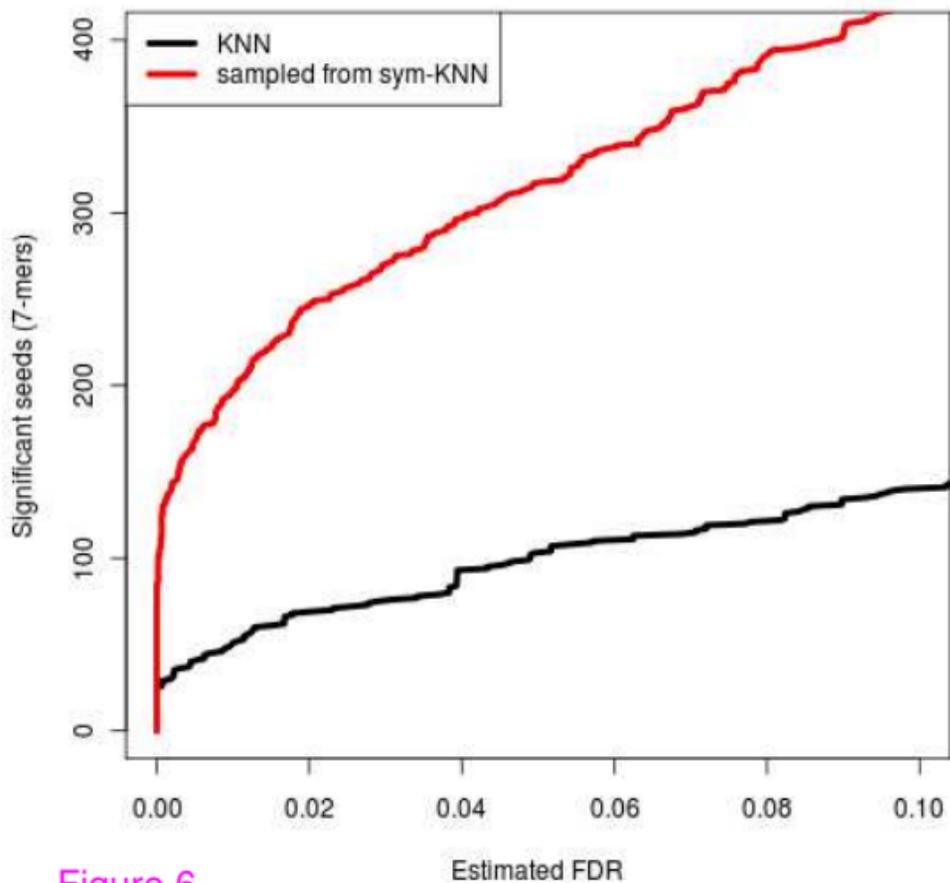


Figure 5

Neighborhood effect on mutual information score



Neighborhood effect on hypergeometric score

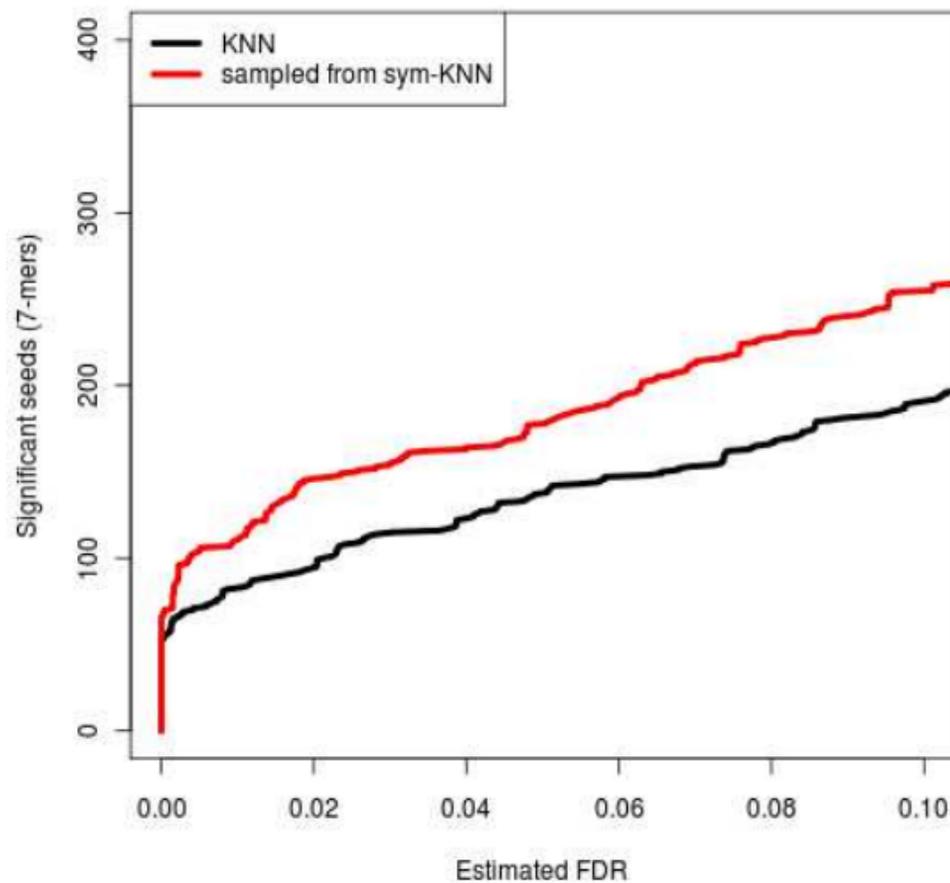


Figure 6

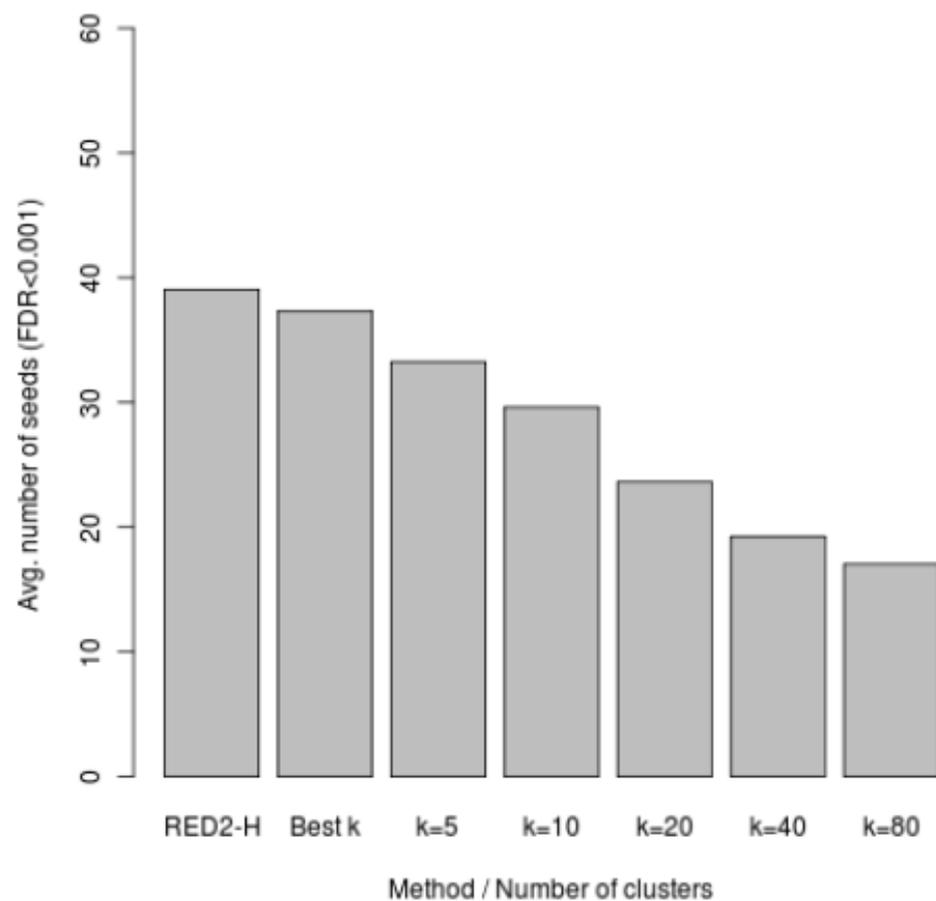
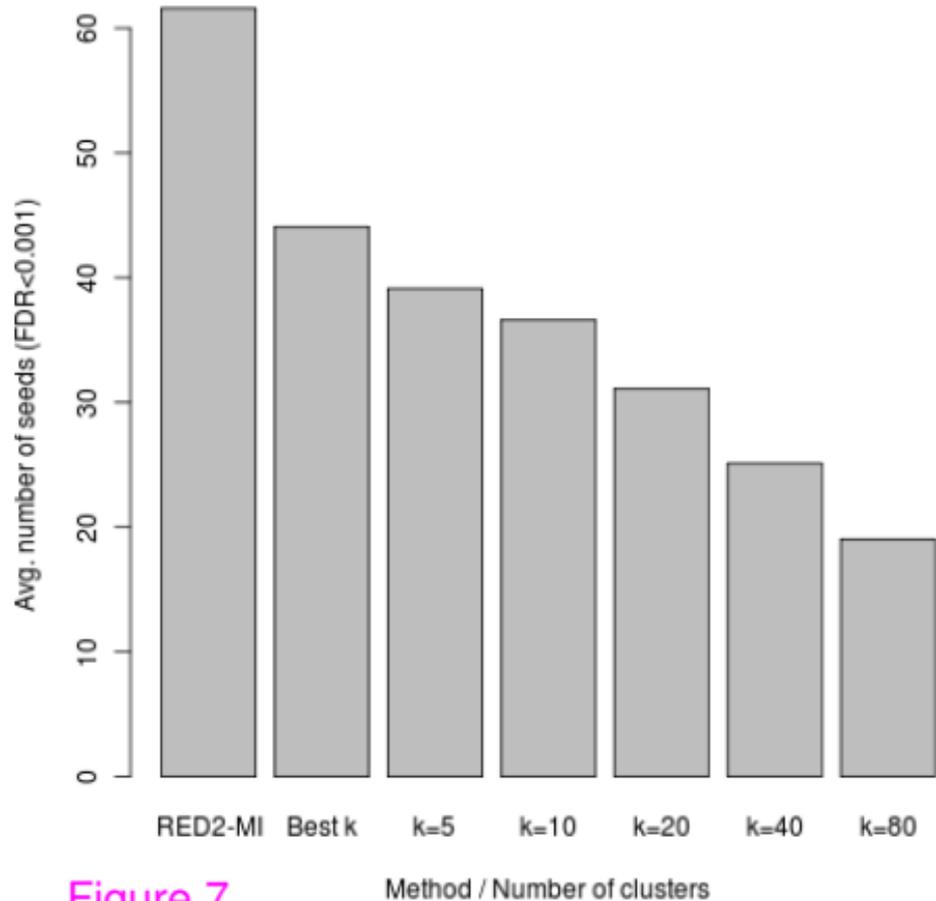


Figure 7

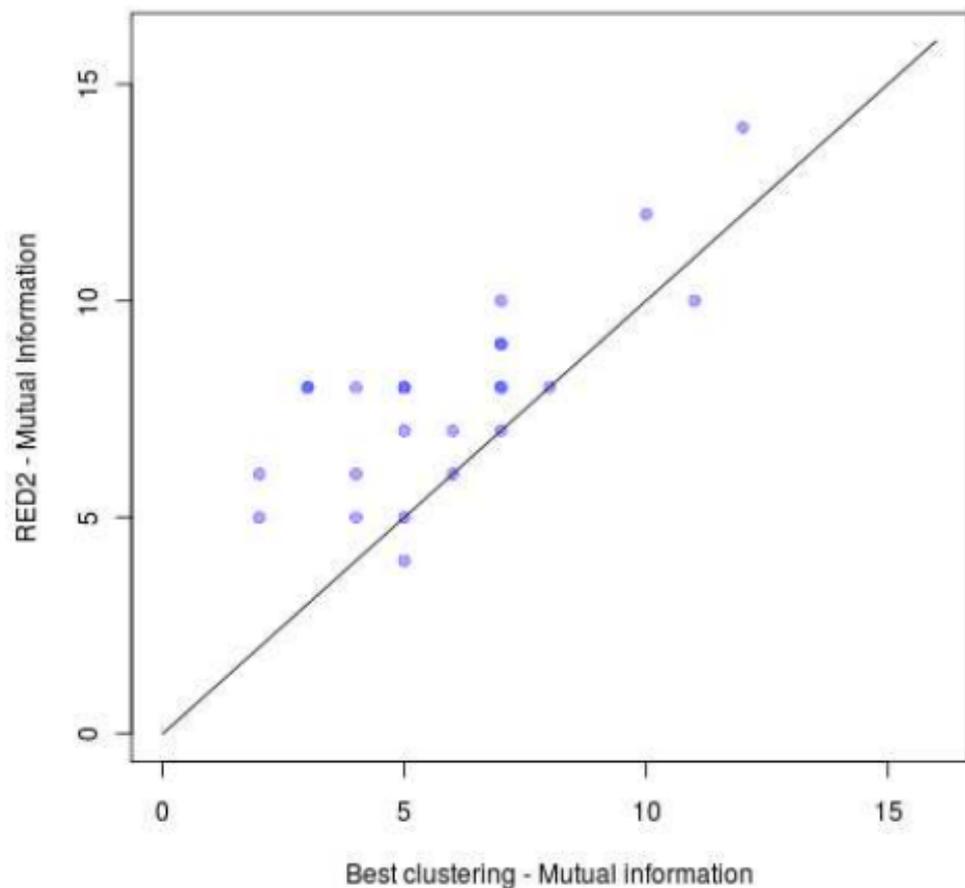
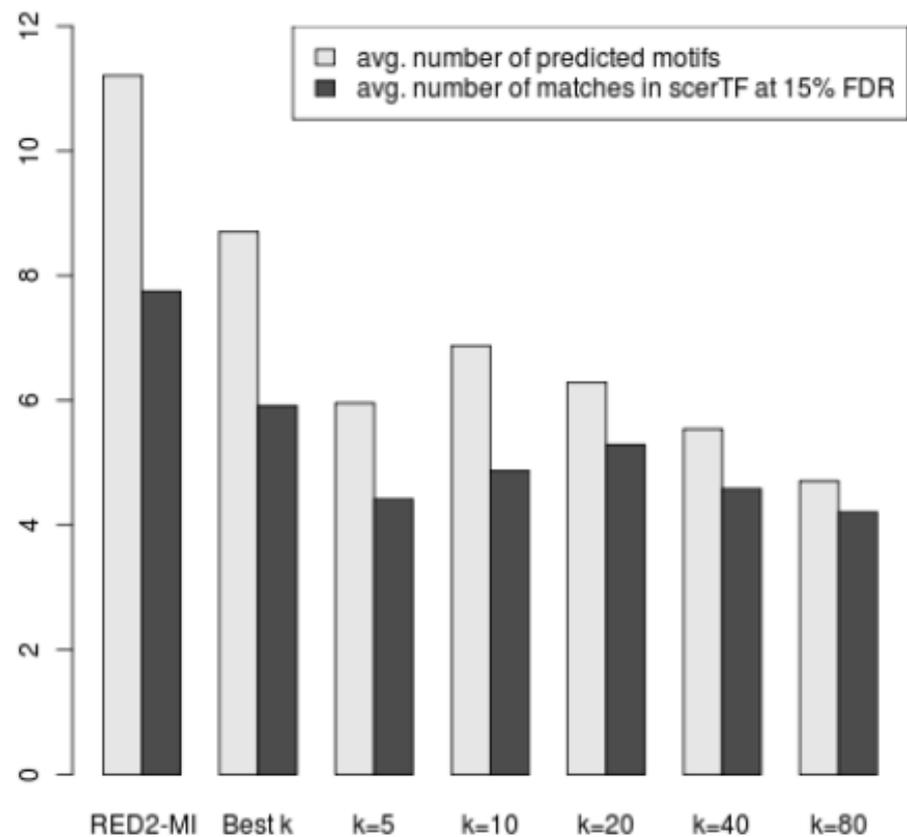


Figure 8

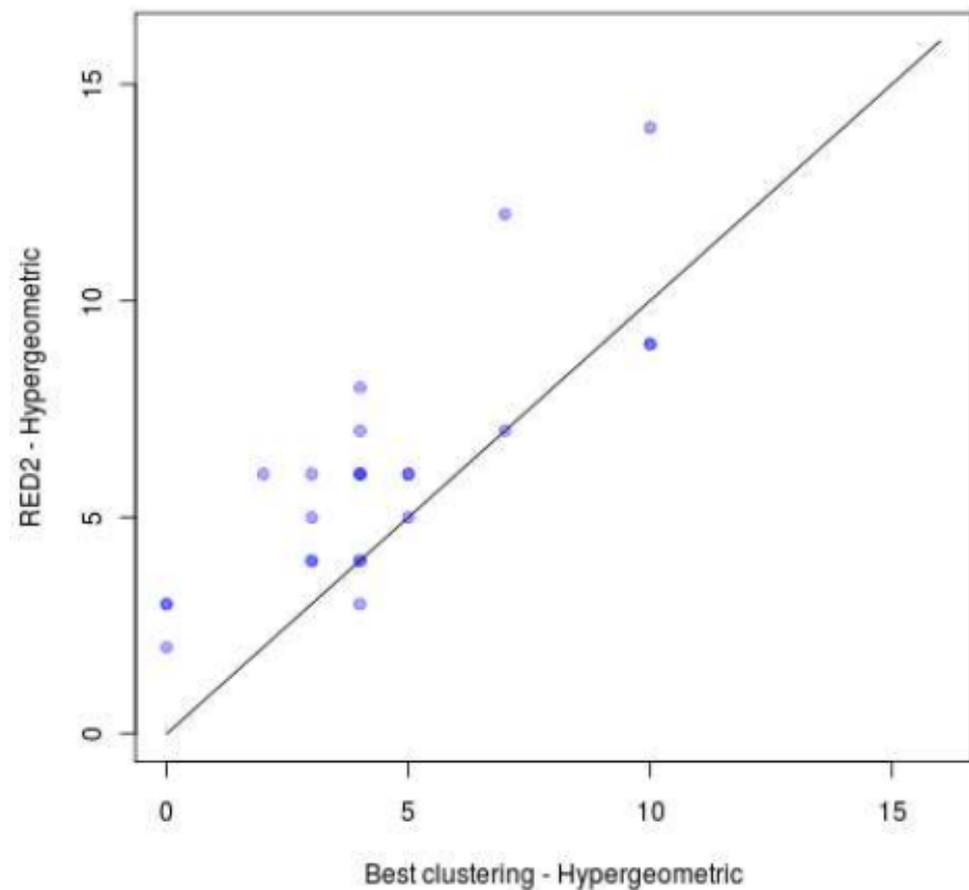
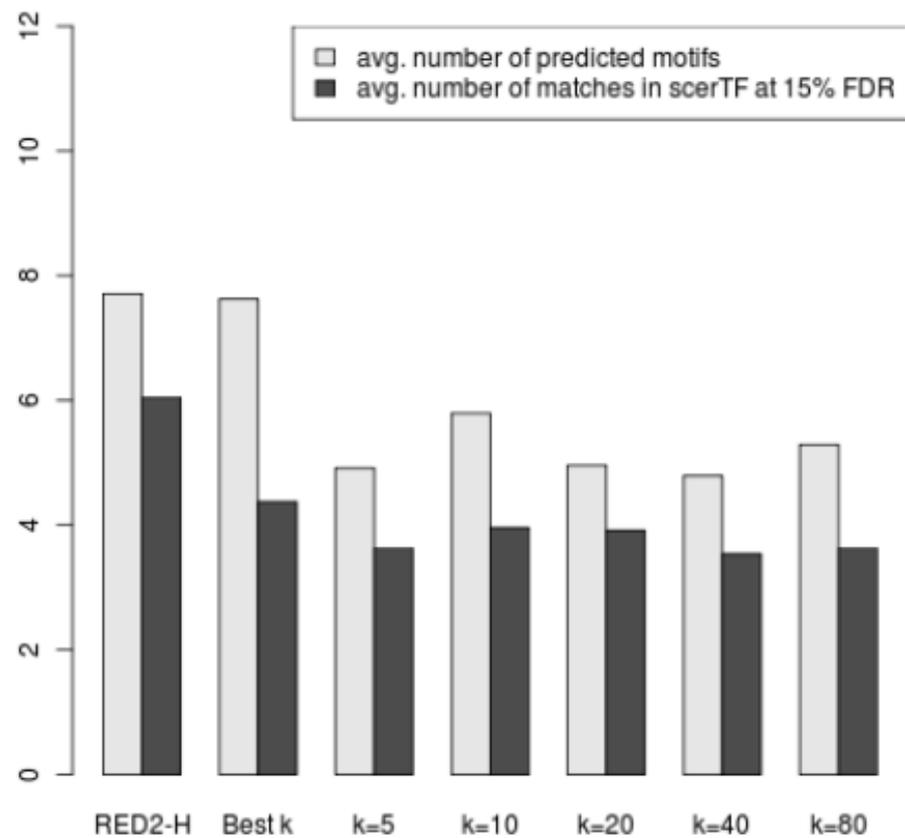


Figure 9

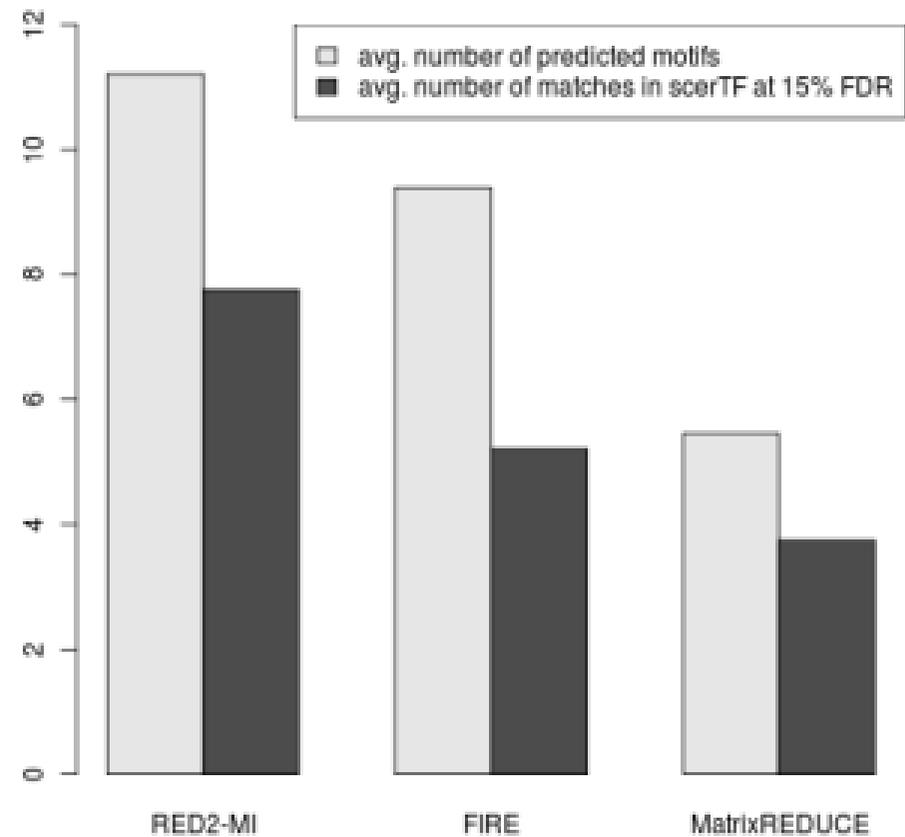
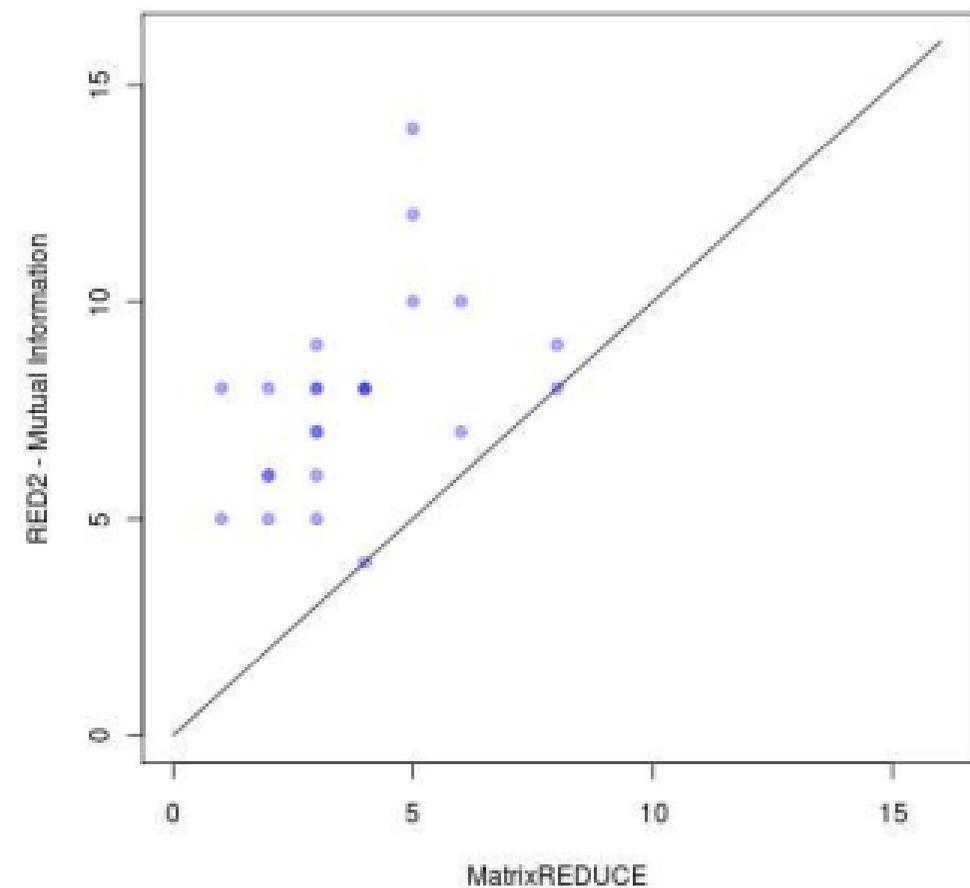
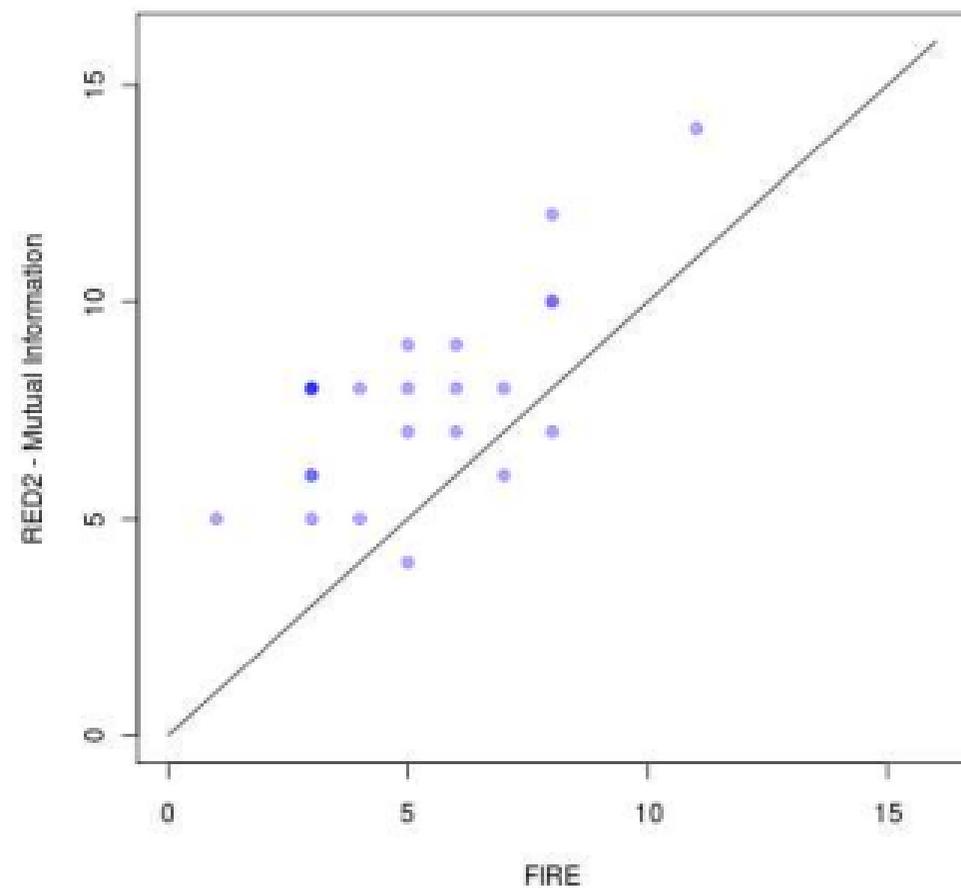
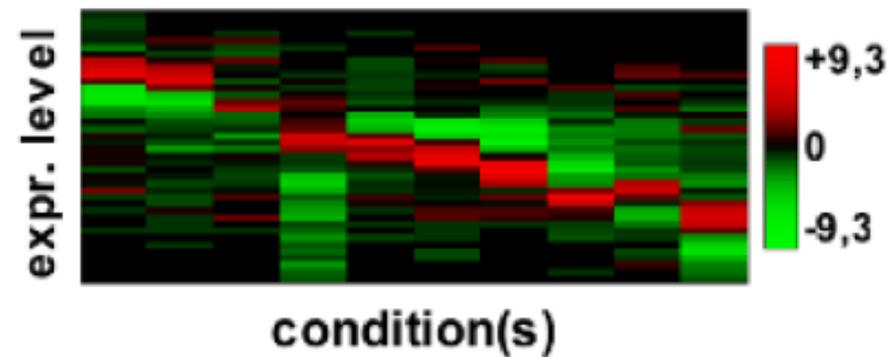
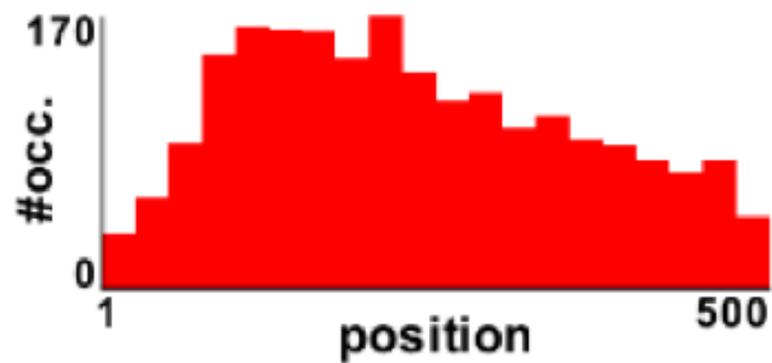
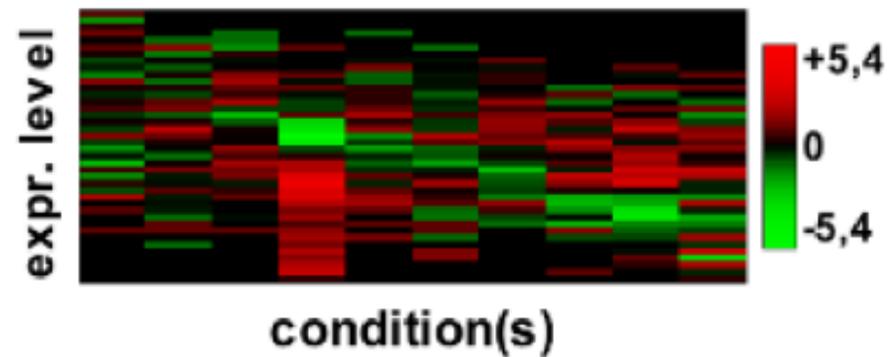
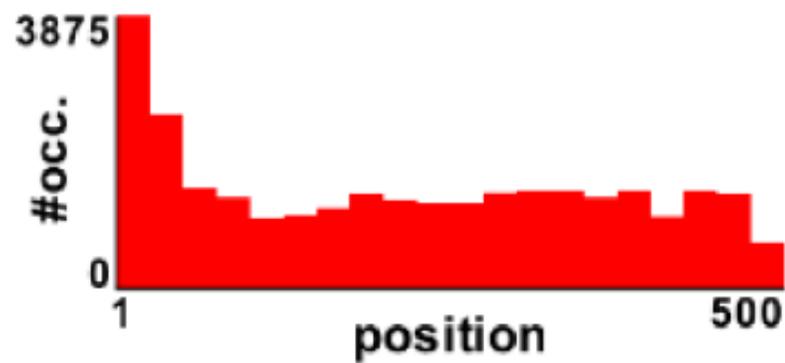


Figure 10





(a)



(b)

Figure 11

### **Additional files provided with this submission:**

Additional file 1: additional\_file\_1\_jaspar\_gordan\_results.pdf, 210K  
<http://genomebiology.com/imedia/1989816179851891/supp1.pdf>

Additional file 2: additional\_file\_2\_gasch\_mi.pdf, 758K  
<http://genomebiology.com/imedia/9863818851891191/supp2.pdf>

Additional file 3: additional\_file\_3\_gasch\_hyper.pdf, 501K  
<http://genomebiology.com/imedia/1802513867851891/supp3.pdf>

Additional file 4: additional\_file\_4\_spellman\_mi.pdf, 535K  
<http://genomebiology.com/imedia/1822381095851892/supp4.pdf>

Additional file 5: additional\_file\_5\_spellman\_hyper.pdf, 537K  
<http://genomebiology.com/imedia/1988688995851892/supp5.pdf>

Additional file 6: additional\_file\_6\_spellman\_fire.pdf, 431K  
<http://genomebiology.com/imedia/9192012858518927/supp6.pdf>

Additional file 7: additional\_file\_7\_spellman\_reduce.pdf, 317K  
<http://genomebiology.com/imedia/1748513560851891/supp7.pdf>

Additional file 8: additional\_file\_8\_recovered\_tf.pdf, 35K  
<http://genomebiology.com/imedia/1928300535851891/supp8.pdf>

Additional file 9: additional\_file\_9\_bozdech\_mi.pdf, 646K  
<http://genomebiology.com/imedia/5915903568518919/supp9.pdf>

Additional file 10: additional\_file\_10\_shock\_mi.pdf, 163K  
<http://genomebiology.com/imedia/4173463685189203/supp10.pdf>

Additional file 11: additional\_file\_11\_bozdech\_fire.pdf, 732K  
<http://genomebiology.com/imedia/1191345654851892/supp11.pdf>

Additional file 12: additional\_file\_12\_bozdech\_reduce.pdf, 279K  
<http://genomebiology.com/imedia/1337545310851893/supp12.pdf>

Additional file 13: additional\_file\_13.txt, 4K  
<http://genomebiology.com/imedia/1670591003851893/supp13.txt>