Predicting gene expression from heterogeneous data.

Matteo Re$^{(1)}$ and Giorgio Valentini$^{(1)}$

(1) Dipartimento di Scienze dell'Informazione,
Università degli Studi di Milano
Via Comelico 39, Italy, {re,valentini}@dsi.unimi.it

Keywords: Weighted averaging, Decision Templates, decision fusion, data integration, gene expression prediction.

Abstract. The complexity of gene expression and the elucidation of the mechanisms involved in its regulation constitute an extremely difficult challenge in modern bioinformatics despite the amount of information made recently available by high-throughput biotechnologies and genome-wide investigations.

In this contribution we investigated the effectiveness of ensemble systems for gene expression prediction. The ability of ensemble systems to integrate heterogeneous datasets allows to exploit not only promoter sequence-based datasets, but also other sources of information, such as phylogenetic patterns of regulatory motifs and covalent histone modifications. To this end we collected data from literature, and we predicted the expression class of 2490 S.Cerevisiae genes using an ensemble of Support Vector Machines trained with 4 different sources of data. The experimental results highlighted that improvement in gene expression prediction performances can be obtained by using ensemble systems. Nevertheless, further investigations are required in order to find the best combination of datasets and data fusion methods for gene-expression class prediction.

1 Introduction

The presence in a living cell of a specific set of transcripts is dynamically regulated in response to variations occurring in its intra and extracellular environment.

A significant part of the ability to regulate gene expression at cellular level is due to the presence of signals encoded in a relatively small region located immediately upstream the Transcription Start Site (TSS), represented by the first nucleotide of the genes, and usually referred to as the core promoter. The classical gene expression regulation models are based on the interactions occurring between the signals encoded in the core promoter region, represented by short oligonucleotide motifs, and a series of proteins collectively named as Transcription Factors (TFs). Only in response of a specific set of environmental conditions, the right combination of TFs bind the short motifs contained in the core promoter region (the Transcription Factor Binding sites, TFBS) and this event enables the cellular transcriptional machinery to start the transcription of the gene.

A key point required for the elucidation of the transcription regulation mechanisms is the definition of the minimal set of information required to detect the presence of specific expression patterns characterizing co-regulated genes.

In a recently published work Tavazoie and colleagues [Beer and Tavazoie, 2004] tried to predict the expression class of S.cerevisiae genes, obtained by clustering gene expression data in many environmental and stress conditions, exploiting only the signals encoded in the core promoter region and achieving a 73%accuracy .

Other sources of information were recently found to be relevant for regulation of gene transcription. Post-translational modifications of histone proteins are able to modulate gene expression patterns [Millar and Grunstein, 2006].

The ChromatinDB database [O'Connor and Wryck, 2007] is dedicated to the storing of information about the statistical over or under-representation of histone modifications at genome-wide level in S-cerevisiae. Other useful information could be obtained by investigating the conservation, at sequence level, of the regulatory motifs located in the promoter regions. In a recent work, MacIsaac and colleagues [McIsaac et al., 2006] published an updated map of conserved regulatory motifs in the yeast genome.

In this contribution we investigate the effectiveness of data integration methods in the expression pattern prediction of 2490 yeast genes using the expression classes identified in [Beer and Tavazoie, 2004]. The experiment was performed using several "late integration" approaches: the classical weighted integration (using two different weighting schemes) and Decision Templates [Kuncheva et al., 2001] in order to provide an overview of the capabilities of multiple classifier systems in the integration of heterogeneous biomolecular data sources for the prediction of gene expression.

To our knowledge, this is the first work devoted to the characterization of performances achievable by data fusion based gene expression prediction using ensemble systems. Our results confirmed the extreme difficulty of the investigated learning task but also clearly indicates that an increment in F-measure, Precision and Recall can be obtained by using data integration methods.

2 Biomolecular data integration with ensemble methods and Decision Templates

2.1 *Reasons for combining biomolecular data through ensembles*

Continuous advances in high-throughput biotechnologies provide new types of data, as well as updates of existing biomolecular data available for gene expression prediction. In this context, ensemble methods are well-suited to embed new types of data or to update existing ones by training only the base learners devoted to the newly added or updated data, without retraining the entire ensemble. This feature of ensemble systems could play a crucial role in gene expression prediction as the definition of the complete list types of data predictive for gene expression is far to be complete. Data fusion of heterogeneous biomolecular data sources can be effectively realized by means of ensemble systems composed by base learners trained on different datasets, and then combining their outputs to compute the consensus decision.

2.2 *Decision Templates and ensembles for gene expression prediction*

In the context of gene expression prediction, as in many other bioinformatics fields, we need to estimate of the reliability of the prediction. To this end, we use SVMs with probabilistic output obtained by applying a sigmoid fitting to their output [Lin et al., 2007]. Thus a trained base classifier computes a function $d_j : X \to [0, 1]$ that estimates the probability that a given example $\mathbf{x} \in X$ belongs to a specific class ω_j. An ensemble combines the outputs of n base learners, each trained on a different type of biomolecular data, using a suitable combining function g to compute the overall probability μ_j for a given class ω_j:

$$\mu_j(\mathbf{x}) = g(d_{1,j}(\mathbf{x}), \ldots, d_{n,j}(\mathbf{x})) \tag{1}$$

A simple way to integrate different biomolecular data sources is represented by the weighted linear combination rule:

$$\mu_j(\mathbf{x}) = \sum_{t=1}^{n} w_t d_{t,j}(\mathbf{x}) \tag{2}$$

The weights are usually computed using an estimate of the overall accuracy of the base learners, but for gene expression prediction, where the expression classes are largely unbalanced (positive examples are largely less than negative ones), we choose the F-measure (the harmonic mean between precision and recall). We consider two different ways to compute the weights:

$$w_t^{\ell} = \frac{F_t}{\sum_{t=1}^{n} F_t} \qquad\qquad w_t^{log} \propto log \frac{F_t}{1 - F_t} \tag{3}$$

The w_t^{ℓ} weights are obtained by a linear combination of the F-measures, and w_t^{log} by a logarithmic transformation. Independently of the choice of the weights the decision $D_j(\mathbf{x})$ of the ensemble about the class ω_j is taken using the estimated probability μ_j (eq. 2):

$$D_j(\mathbf{x}) = \begin{cases} 1, & \text{if } \mu_j(\mathbf{x}) > 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where output 1 correspond to positive predictions for ω_j and 0 to negatives.

Certain types of biomolecular data can be informative for some expression classes, but uninformative for others. Hence it would be helpful to take into account whether certain types can be informative or not, depending on the class to be classified. To this end *Decision Templates* [Kuncheva et al., 2001] can represent a valuable approach.

More precisely, the decision profile DP(**x**) for an instance **x** is a matrix composed by the $d_{t,j}$ ∈[0,1] elements representing the support given by the t^{th} classifier to class ω_j. Decision templates DT_j are the averaged decision profiles obtained from $\mathbf{X}_j$, the set of training instances belonging to the class ω_j:

$$DT_j = \frac{1}{|\mathbf{X}_j|} \sum_{\mathbf{x} \in \mathbf{X}_j} DP(\mathbf{x}) \tag{5}$$

Given a test instance we first compute its decision profile and then we calculate the similarity $\mathcal{S}$ between $DP(\mathbf{x})$ and the decision template DT_j for each class ω_j, from a set of c classes. As similarity measure the Euclidean distance is usually applied:

$$\mathcal{S}_j(\mathbf{x}) = 1 - \frac{1}{n \times c} \sum_{t=1}^{n} \sum_{k=1}^{c} [DT_j(t,k) - d_{t,k}(\mathbf{x})]^2 \tag{6}$$

The final decision of the ensemble is taken by assigning a test instance to a class with the largest similarity:

$$D(\mathbf{x}) = \arg\max_j \mathcal{S}_j(\mathbf{x}) \tag{7}$$

In our experimental setting we consider dichotomic problems, because a gene may belong or not to a given expression class, thus obtaining two-columns decision template matrices.

It is easy to see that with dichotomic problems the similarity ($\mathcal{S}_1$) (eq. 6) for the positive class and the similarity ($\mathcal{S}_2$) for the negative class become:

$$\mathcal{S}_1(\mathbf{x}) \;=\; 1 - \frac{1}{n} \sum_{t=1}^{n} [DT_1(t,1) - d_{t,1}(\mathbf{x})]^2 \tag{8}$$

$$\mathcal{S}_2(\mathbf{x}) \;=\; 1 - \frac{1}{n} \sum_{t=1}^{n} [DT_2(t,1) - d_{t,1}(\mathbf{x})]^2 \tag{9}$$

where DT_1 is the decision template for the positive class and DT_2 for the negative one. The final decision of the ensemble for a given gene expression class is:

$$D(\mathbf{x}) = \arg\max_{\{1,2\}} (\mathcal{S}_1(\mathbf{x}), \mathcal{S}_2(\mathbf{x})) \tag{10}$$

Table 1: Datasets

Code	Dataset	examples	features	description
D_{tavR}	Beer motif scores real	2587	666	Beer motif scores (Real) from [Beer and Tavazoie, 2004]
D_{tavB}	Beer motif scores binary	2587	666	Beer motif scores (binary) from [Beer and Tavazoie, 2004]
$D_{histmod}$	Histone modification scores	2580	22	Histone modification scores collected from the ChromatinDB [O'Connor and Wryck, 2007] database
D_{phylo}	Motifs conservation scores	2492	121	Motifs conservation scores produced using the PhyloCon algorithm [McIsaac et al., 2006]

3 Experimental setup

We chose to perform our experiments starting from the S. cerevisiae data provided in [Beer and Tavazoie, 2004].

We also included two additional datasets collected, respectively, from the ChromatinDB database [O'Connor and Wryck, 2007] and from [McIsaac et al., 2006] supplemental material.

The motifs scores used as indicators of the presence/absence of the TFBSs in the gene promoters in [Beer and Tavazoie, 2004] were used in the form provided by the authors and in form of binary indicators.
Genome-wide Chromatine Immuno Precipitation (ChIP) data for 22 different histone modifications were downloaded from ChromatinDB [O'Connor and Wryck, 2007]. We extracted from ChromatinDB all the available data inherent to ChIP data annotated in the genomic regions corresponding to all the annotate S.cerevisiae gene promoters. The last dataset involved in our experiments is based on the conservation scores produced by the PhyloCon algorithm [McIsaac et al., 2006]. The authors provided these data in form of three tables of motifs scores expressing the conservation level of the motifs annotated in S.cerevisiae promoters produced by comparative genomics methods based on the comparison of orthologous promoters pairs. The three tables refer to low, moderately and highly conserved motifs. The PhyloCon data were merged into an unique table expressing the conservation level of all the TFBSs in form of discrete and ordered indicators ranging from 0 (not conserved) to 3 (highly conserved).

The main characteristics of the data sets used in the experiments are summarized in Tab. 1.

We considered yeast genes common to all data sets (2490), and we associated them to the expression classes reported in [Beer and Tavazoie, 2004]. The investigated classification problems are affected by a severe unbalance between positives and negatives examples: the number of positive examples is between 5.0% and 0.5% of the available data depending on the considered expression class. In order to avoid classification tasks with a too low number of positive examples the 8 smallest expression classes were excluded from our experiments resulting into a 41 classification problems. The learning problem was split in 41 binary classification tasks in which each gene was predicted as belonging or not to the considered expression class.
Each dataset was split into a training set and a test set (composed,respectively, by the 70% and 30% of the available samples). We performed a 3-fold stratified cross-validation on the training data for model selection: we computed the F-measure across folds, while varying the parameters of gaussian kernels (both σ, ranging from 10^-5 to 10^5, and the C regularization term, ranging from 10^-5 to 10^5).
Classification performances of the component classifiers and the ensemble systems have been evaluated using a multiple hold-out scheme based on 5 replicates of the aforemen-

Table 2: Balanced setup. Ensembles of learning machines, average performances of base learners and performances of D_{tavR}: average F-measure, accuracy, precision and recall computed by multiple hold-out techniques.

Metric	E_{lin}	E_{log}	E_{dt}	D_{avg}	D_{tavR}
F	0.789	0.789	**0.791**	0.699	0.783
acc	0.791	0.791	0.793	0.660	0.785
prec	0.799	0.799	0.800	0.649	0.789
rec	0.790	0.790	0.790	0.813	0.7910

tioned training and testing procedure. The collected test sets classification performances have been averaged across all the replicates.
In order to evaluate the gain in prediction performances achievable by data integration methods in presence and absence of the problems due to the unbalance between positives and negatives examples we repeated the entire procedure using artificially balanced datasets constituted by all the positive examples belonging to the considered expression class and the same amount of negative examples randomly chosen from the remaining expression classes.

The just described experimental setup resulted into 41 $\times$ 7 $\times$ 5 $\times$ 2 = 2870 pairwise classification tasks.

We adopted many performances evaluators, instead of the Accuracy used by Beer and colleagues [Beer and Tavazoie, 2004]. Our choice is motivated by the large unbalance between positive and negative examples that characterizes the investigated prediction problems: indeed on the average only a small subset of the available genes is annotated to each expression class. We compared the performances of single gaussian SVMs trained on each data set with those obtained with the ensembles described in Sect. 2.2. We normalized the data with respect to the mean and standard deviation, separately for each data set.

4 Results

The summary of the averaged results collected in the artificially balanced gene expression prediction tasks are reported in Tab. 2. The table shows the average F-measure,accuracy, precision and recall across the 41 selected gene expression classes, obtained through the evaluation of the test sets (each constituted by 747 genes).The performances are estimated using a multiple hold-out based on 5 replicates and the final test sets performances are averaged. The three first columns refer respectively to the weighted linear, logarithmic linear and decision template ensembles (see Sect. 3), D_{avg} represents the averaged results of the single SVMs across the four datasets, and D_{tavR} represents the single SVM trained using data provided by Tavazoie and colleagues (Tab. 1). Tab. 3 shows the same results obtained in the unbalanced learning tasks.

Table 3: Unbalanced setup. Ensembles of learning machines, average performances of base learners and performances of D_{tavR}: average F-measure, accuracy, precision and recall computed by multiple hold-out.

Metric	E_{lin}	E_{log}	E_{dt}	D_{avg}	D_{tavR}
F	0.109	0.142	**0.268**	0.108	0.209
acc	0.923	0.948	0.912	0.977	0.977
prec	0.244	0.333	0.436	0.216	0.429
rec	0.141	0.137	0.256	0.080	0.156

Looking at the values presented in Tab. 2 and considering the F-measure, we see that in the artificially balanced setup, on the average data integration through ensemble methods provides better results than single SVMs, independently of the applied combination rule. In particular Decision Templates achieved the best average F-measure albeit the performances are quite similar for all the tested combination methods. Considering the averaged accuracies the data fusion methods are still able to outperform all the component SVMs. The observed trend is confirmed for Precision but not for the Recall: only the Decision Templates combiner was able to outperform all the component classifiers independently of the considered performance metric.

The performances obtained by the component classifiers in the 41 separated expression class prediction tasks highlighted that the performances of the classifiers trained on the $D_{histmod}$ and D_{phylo} are, on the average, lower than the ones obtained by the classifiers trained using the matching scores used in [Beer and Tavazoie, 2004], even if they were able to outperform the classifiers trained using the [Beer and Tavazoie, 2004] data in some expression classes prediction tasks.
Under this balanced setup, and using the accuracy as performance metric, we outperformed the results obtained by Tavazoie and colleagues (73%).
Under the unbalanced experimental setup (see Tab. 3), the large accuracies are due to the concurrent failure of the component classifiers in the learning problems (meaning that in many classification tasks all the test instances were predicted as negatives), and the large unbalance in the data. They cannot thus be used to evaluate the performances of classifier systems. According to the collected F-measures we still observe an improvement in performance achievable using the ensemble systems but with a different pattern than the one emerging from the results achieved under the balanced setup. In particular the linear combiners (E_{lin} and E_{log}) are on the average unable to outperform the best performing base learner (D_{tavR}). The only combiner able to outperform the best component classifier (in 33 over 41 classification tasks) is the Decision Template combination rule. The ability of E_{dt} to outperform D_{tavR} is also confirmed looking at the Precision and the Recall.

In this extremely difficult classification task the results confirmed the ability of the Decision Templates combiner to learn not only from correct predictions but also from the wrong ones exploiting the different patterns in the errors produced during the classification of the positive and negative instances. According to the collected F-measures averaged for each gene expression class across the performed replicates, E_{lin},E_{log} and E_{dt} were able to outperform the best component classifier (D_{tavR}) respectively 24, 23 and 27 times under the balanced setup and 4,2 and 33 times under the unbalanced setup indicating that in critically difficult gene expression prediction problems the Decision Templates ensemble system is the safer choice. The averaged F-measure performances of the different methods, under the artificially balanced setup are summarized in Fig. 1: the ensemble system performances are quite similar to those obtained by the D_{tavR} SVM.

The same detailed performances information collected under the unbalanced setup are reported in Fig. 2: the SVMs and all the ensembles (data not shown) predicted any example as negative producing F-measures equal to 0 in 7 out of 41 learning tasks (expression classes 6,13,22,23,31,34 and 39). In the remaining learning tasks E_{dt} was able, on the average, to outperform D_{tavR}. The ensemble system was able to provide better performances in learning tasks in which the averaged performances of the component classifiers were close to 0 (see classes 17,18 and 25). It is worth noting that, under the unbalanced experimental setup, all the tested component classifiers and ensemble systems failed to learn the separation between positive and negative examples, resulting in a final F-measure of 0 in 7 out of 41 learning tasks. In [Chin et al., 2005] the authors investigated the distribution, at genome-wide level, of the evolutionary constraints in the

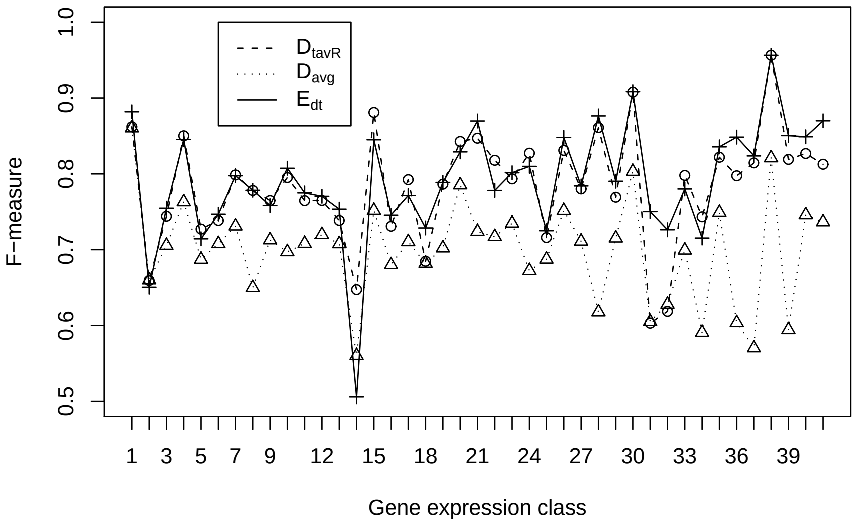

Figure 1: Comparison of the F-measures achieved in gene expression prediction: D_{avg} stands for the average across SVM single learners, D_{tavR} for the best single SVM, E_{dt} for decision template ensemble.

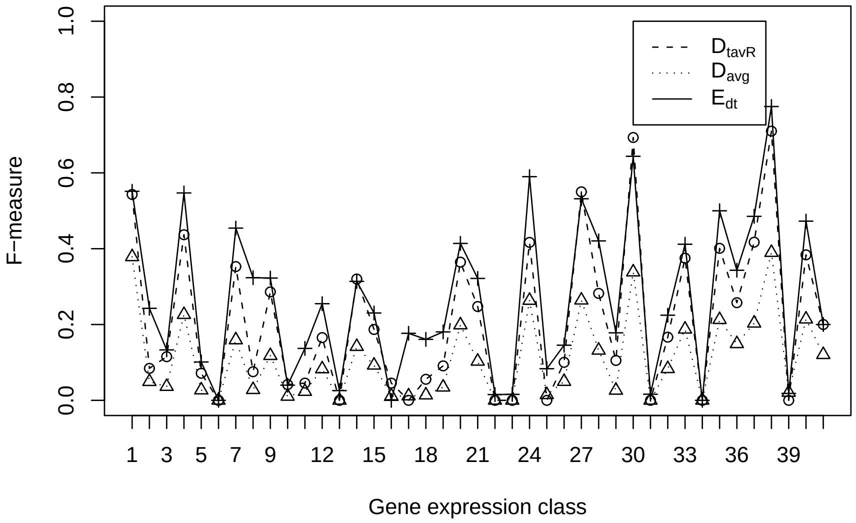

Figure 2: Comparison of the F-measures achieved in gene expression prediction: D_{avg} stands for the average across SVM single learners, D_{tavR} for the best single SVM, E_{dt} for decision template ensemble.

S.cerevisiae genome in the aim to separate functionally conserved and neutral sequences in the promoters. The authors also investigated the eventual existence of an association between the functional class of the genes and the length of the high conserved regions (HCRs) detectable in their promoters. The authors observed that the promoters of genes belonging to GO terms enriched in constitutively expressed (housekeeping) genes are characterized by the presence of shorter HCRs if compared to promoters of highly regulated genes. A manual inspection of the genes contained in the expression clusters of the 7 expression classes in which we failed to obtain a F-measure greater than 0 highligted an enrichment in housekeeping genes (like aminoacyl-tRNA-synthetases and ORFs involved in the transcription of rRNAs). Even if the lack of evolutionary pressure detectable in the housekeeping genes promoters can't be directly related with the specificity and the regulatory strength of the motifs present in their sequences, these promoters are not, by definition, expected to contain patterns of regulatory motifs able to drive a fine regulated (and thus well defined) expression pattern.

5 Conclusions

In this work we investigated the impact on yeast gene expression prediction performances of ensemble-based data fusion methods. Our experiments demonstrated the potential benefits introduced by the usage of simple ensemble-based prediction systems for the integration of multiple sources of data in gene expression classification problems. Despite the extreme difficulty of the investigated classification problems, ensemble systems achieved good performances if compared with the best performing component classifier trained on the matching motif scores used in [Beer and Tavazoie, 2004]. Once removed the large unbalance in the data, sampling randomly an amount of negative examples equal to the number of the positive ones, we obtained better averaged performances than those reported by Tavazoie and colleagues [Beer and Tavazoie, 2004]. The ability of ensemble systems to exploit the diversity of the component classifiers predictions in order to improve the classification performances is more apparent in the unbalanced classification tasks, representing a more realistic view of real world problems. We think that the application and the development of more refined ensemble methods, exploiting the modularity and scalability that characterizes the ensemble approach, represent a promising research line for gene expression prediction using heterogeneous sources of complex biomolecular data.

The choice of data sources potentially informative for the prediction of expression patterns from sequence data is still an open problem but, according to very recent findings pointing out that bidirectional promoters are responsible for pervasive transcription in the S.cerevisiae genome [Xu et al., 2009] it might be of great interest the usage of features obtained not only from the sequence located upstream but also downstream the TSS. In particular, the usage of epigenetic and phylogenetic patterns could help in the elucidation of the complex mechanisms underlying the fine regulation of genes transcription.

Acknowledgments

The authors would like to gratefully acknowledge partial support by the PASCAL2 Network of Excellence under EC grant no. 216886. This publication only reflects the authors' views.

References

[Beer and Tavazoie, 2004] Beer, M. and Tavazoie, S. (2004). Predicting gene expression from sequence. *Cell*, 117.

[Chin et al., 2005] Chin, C., Chuang, J., and Li, H. (2005). Genome wide regulatory complexity in yeast promoters: seperation of functionally conserved and neutral sequence. *Genome research*, 15.

[Kuncheva et al., 2001] Kuncheva, L., Bezdek, J., and Duin, R. (2001). Decision templates for multiple classifier fusion: an experimental comparison. *Pattern Recognition*, 34(2):299–314.

[Lin et al., 2007] Lin, H., Lin, C., and Weng, R. (2007). A note on Platt's probabilistic outputs for support vector machines. *Machine Learning*, 68:267–276.

[McIsaac et al., 2006] McIsaac, K., Wang, T., Gordoni, B., Gifford, D., Stormo, G., and Fraenkel, E. (2006). An improved map of conserved regulatory sites map for saccharomyces cerevisiae. *BMC Bioinf.*, 7.

[Millar and Grunstein, 2006] Millar, C. and Grunstein, M. (2006). Genome-wide patterns of histone modifications in yeast. *Nat. rev. Mol. Cell. Biol.*, 7.

[O'Connor and Wryck, 2007] O'Connor, T. and Wryck, J. (2007). Chromatindb: a database of genome-wide histone modification patterns for saccharomyces cerevisiae. *Bioinformatics*, 23.

[Xu et al., 2009] Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Munster, S., Camblong, J., Guffanti, E., Stutz, F., Huber, W., and Steinmetz, L. (2009). Bidirectional promoters generate pervasive transcription in yeast. *Nature*, 457.