Assessment of algorithms for inferring positional weight matrix motifs of transcription factor binding sites using protein binding microarray data

Yaron Orenstein[1], Chaim Linhart[1] and Ron Shamir[1,*]

[1] Blavatnik School of Computer Science, Tel-Aviv University, Tel-Aviv, 69978, Israel

* To whom correspondence should be addressed. Tel: +972-3-640-5383; Fax: +972-3-640-5384; Email: rshamir@tau.ac.il

Present Address: Ron Shamir, Blavatnik School of Computer Science, Tel-Aviv University, Tel-Aviv, 69978, Israel

ABSTRACT

The new technology of protein binding microarrays (PBMs) allows simultaneous measurement of the binding intensities of a transcription factor to tens of thousands of synthetic double-stranded DNA probes, covering all possible 10-mers. A key computational challenge is inferring the binding motif from these data. We present a systematic comparison of four methods developed specifically for reconstructing a binding site motif represented as a positional weight matrix from PBM data. The reconstructed motifs were evaluated in terms of three criteria: concordance with reference motifs from the literature and ability to predict *in vivo* and *in vitro* bindings. The evaluation encompassed over 200 transcription factors and some 300 assays. The results show a tradeoff between how the methods perform according to the different criteria, and a dichotomy of method types. Algorithms that construct motifs with low information content predict PBM probe ranking more faithfully, while methods that produce highly informative motifs match reference motifs better. Interestingly, in predicting high-affinity binding, all methods give far poorer results for *in vivo* assays compared to *in vitro* assays.

INTRODUCTION

Understanding gene regulation is a fundamental problem in biological research. A principal way to regulate gene expression in the cell is via transcription, which is governed primarily by transcription factors (TFs). A TF is a protein that binds to the promoter region of a gene at specific sequences, called TF binding sites (TFBSs). The binding of one or several TFs enables or impedes the transcription of the gene. A TF binds to similar short nucleotide sequences at different affinities. Finding these cis-regulatory elements and modeling the affinity of TF binding to them is a central challenge in understanding gene regulation.

The most common computational model for describing a TFBS motif is a position weight matrix (PWM) [1]. The TFBS is represented by a 4×*k* matrix, where *k* is the motif length. Each column contains four probabilities, representing the nucleotide frequencies at that position. This relatively simple model is highly popular since it is compact, effective and easy to interpret.

New technologies have enabled comprehensive mapping of protein-DNA binding affinities. The main technology to measure *in vivo* protein occupancy is chromatin immunoprecipitation (ChIP). In the ChIP-chip method, the protein-bound DNA segments are hybridized to a pre-designed microarray [2], whereas the ChIP-seq method uses deep sequencing to read the bound DNA segments [3]. A recent promising technology in this field is the protein binding microarray (PBM) [4]. This microarray contains ~41,000 synthesized, 60bp-long double-stranded DNA probes, each containing 36bp of unique sequence, designed so that every possible 10-mer is contained in exactly one probe sequence. A single *in vitro* experiment measures the binding intensity profile of a specific TF to each probe, thereby providing complete coverage of the binding affinity of the TF to all possible 10-mers. Often, two experiments with different array designs are performed with the same TF, providing *paired* profiles.

Numerous computational methods for finding a motif in a target set of promoters have been developed over the last two decades [5-7]. Predicting binding sites based on PBM data is different: the experimental data are much more comprehensive, covering all possible 10-mers, but are generated *in vitro* and in a high-throughput (and hence noisy) fashion. Therefore, several methods were recently developed specifically for identifying TFBS motifs from PBM profiles. Here we compare methods that represent the motifs as PWMs. We do not include methods that use more complex models [8], since we choose to focus on simpler, more compact models.

In this paper we present a systematic comparison of four algorithms for identifying TFBS motifs from PBM profiles: Seed-and-Wobble (SW) [4], RankMotif++ (RM) [9], BEEML-PBM (BE) [10] and the algorithm Amadeus-PBM (AM) introduced here (see **Table 1**). In 2005, a systematic comparison of computational methods for motif discovery in promoters clarified some of the issues and the difficulties in that domain, and led to progress in that research area [11]. We hope that our study will have a similar effect regarding methods for analyzing PBM data.

RESULTS

Concordance with SELEX-based reference motifs from the literature:

We used each method to find motifs using PBM data, and compared the results to previously reported motifs for the same TFs, obtained using independent experiments. Each motif was learned using the data from two paired experiments performed with the same TF. For each TF, we measured the distance between the PBM-based PWM to the PWM of the same TF as published in JASPAR [12]. For this test we used all mouse PBM datasets from the SCI09 study [13,14] that had a corresponding PWM in JASPAR, excluding those for which the JASPAR PWMs were constructed using PBM data. This set contained 58 PWMs. Most were constructed based on *in vitro* SELEX experiments, which are still the main source of TF motifs.

The AM PWMs were the most similar to JASPAR, with average Euclidean distance (± estimated standard deviation) 0.178±0.11. The average for SW was 0.193±0.1, for RM was 0.21±0.09, and for BE was 0.227±0.1 (**Table 3**). The difference between AM and SW was not significant (p=0.17, Wilcoxon rank-sum test) and both were significantly better than RM and BE (p=0.001 and p=0.0005 compared to AM, respectively).

We then focused on high-quality predictions of the four methods. We say that a motif is successfully recovered by a method if the Euclidean distance of the predicted PWM from the reference PWM is below a predetermined cutoff. As in [15], we used three cutoffs for the distance. AM attained a higher success rate using all cutoffs (**Figure 1**). A similar comparison of mouse motifs in TRANSFAC [16] and yeast motifs in ScerTF [17], and a parallel comparison, using p-value for the significance of the similarity [18], showed a similar advantage to AM (**Figure S1**).

Visual inspection suggested that the PWMs produced by AM and SW are easier to interpret and look distinct in logo format (**Figure 2**). To quantify this observation, we calculated the average information content for each PWM (see **Supplementary Methods**). Averaged over the PWMs computed from all 115 available paired mouse PBM sets, the information scores for the raw PWMs were 1.03, 0.61, 0.42 and 0.53 bits for AM, SW, RM and BE, respectively, with AM scoring significantly higher ($p<10^{-15}$, Wilcoxon rank-sum test). After trimming the PWMs to discard flanking positions with low information, the information averages were 1.03, 1.09, 0.54 and 0.61 bits, respectively ($p=1.2\cdot10^{-7}$ when comparing SW to AM and $<10^{-15}$ when comparing AM and SW to RM and BE). The full comparison results are available in **Table S1**.

Predicting *in vitro* binding intensities:

Next, we tested the prediction of binding intensities by the four methods on 115 pairs of mouse PBM profiles [13,14] following the procedure in [9]. Each method learned a PWM according to one PBM experiment; this PWM was used to rank the probes of its paired array. The goal was to correctly rank the positive probes, i.e. those with highest affinity measurements. The set of positive probes (denoted 4σ, see **Supplementary Methods**) contained an average of 912 probes per array. We also evaluated larger sets of positive probes using more permissive cutoffs (denoted 3σ, 2σ and 1σ; an average of 1580, 3215 and 8224 probes per array, respectively).

When testing on 4σ top probes set (**Table 3** and **Figure 3**), BE had significantly best Spearman and AUC scores (p<0.0025, Wilcoxon rank-sum test), while AM and RM were essentially equal (p=0.41 and p=0.44, respectively), and significantly better than SW ($p<10^{-4}$). Using the sensitivity measure, BE was again best ($p<10^{-15}$), AM second best ($p=3.8\cdot10^{-6}$ compared to SW), and RM and SW were roughly the same (p=0.18). Hence, BE showed consistently best performance in all three measures, followed by AM. Interestingly, BE gave the poorest AUC and Spearman scores on a few samples. On larger probe sets (**Figure 4**), BE performed best, followed by RM. The AUC and sensitivity criteria deteriorated for all methods, as expected due to the increasing difficulty in ranking lower-affinity probes. The Spearman score improvement results from its bias to larger sets, so it is more meaningful for comparison of sets of similar sizes. Full results are available in **Table S2**.

Predicting *in vivo* binding intensities:

Since PBM and SELEX are *in vitro* assays, which may introduce biases, we also tested the methods' abilities to predict binding intensities for *in vivo* experiments. Our evaluation included ChIP-chip datasets of 32 yeast TFs (69 experiments) that had also PBM profiles [8,19]. A PWM learned according to the profiles of both PBMs (when available) is tested against the data from a ChIP-chip experiment. To evaluate the prediction on the high intensity promoters, where binding is expected to be strongest, we used the positive promoter set as those with reported p-values below 0.001.

All methods performed quite similarly on the AUC and Spearman rank coefficient criteria (**Table 2**). Using the sensitivity measure, SW was better than the other three (p<0.02), AM and BE were roughly the same (p=0.39) and significantly better than RM (p<0.04). Hence, SW showed consistently best performance in all three measures, while AM and BE were second best (**Table 3** and **Figure 4**). Full results are available in **Table S3**.

Running times:

We ran each method on the same 10 examples using a single core of an Intel® Xeon® CPU E5410 @ 2.33GHz, with 6MB of cache and 16GB of memory. On average, AM runs for 30 seconds (including pre-processing), while BE, RM and SW run for about 15 minutes, one hour and more than two hours, respectively (**Table 3**). BE currently uses SW results as seeds, thus SW's running time should be added to the total running time of BE. Hence, AM provides a speedup by a factor of 30–200.

Similarity between the algorithms:

We evaluated the similarity between the PWMs produced by the four algorithms (**Table 4A**). In terms of PWM distance, the pairs AM/SW and RM/BE were more similar than others. Note that the comparison is not symmetrical, since it uses the eight most informative contiguous positions in the first PWM (corresponding to a column in the table). Large asymmetries (e.g., SW-RM and RM-SW) reflect the fact that these positions are not clearly detectable in RM and BE PWMs (see also **Figure 2**). On average, the distance between PWMs from different methods is similar to the distance between these and the reference PWMs (**Table 3**).

We also compared the probe ranking that the PWMs of the different algorithms induce (**Table 4B-D**). We used a PWM inferred by one algorithm on a PBM to rank the probe set of the paired PBM, and measured sensitivity and AUC for these probes ranking produced by another algorithm. Results tended to show more symmetry, with pairs involving BE obtaining best scores, in agreement with the good performance of BE in ranking (**Figure 3** and **Table 3**). Additionally, we focused on rankings of the 4σ probe set and compared them using Spearman rank coefficient. PWMs inferred by two algorithms on a PBM to rank the 4σ probe set of the paired PBM, and compared the two rankings using Spearman score. Again pairs with BE got the highest scores, and remarkably, all pair scores were much higher than their similarity scores to original binding intensities (Spearman rank coefficient, 0.5-0.6 compared to 0.24-0.31, respectively).

DISCUSSION

We have described an assessment of four tools for extracting binding site motifs from PBM data. All four methods report their results in the form of a positional weight matrix (PWM). **Table 3** summarizes the comparison. All tools were run with their recommended default parameters; tuning the parameters could improve the results of some methods and affect the relative ranking in our test criteria.

The reference motifs stored in databases are strongly dependent on experimental sources. Most TRANSFAC and JASPAR motifs that we used were created based on SELEX, an *in vitro* assay of limited accuracy and throughput. Still, the relative performance of the methods was essentially the same when tested on three different databases of two species, which indicates robustness of our conclusions.

The best results in similarity of reference mouse motifs to predicted motifs from PBMs (**Figure 1**) were comparable to the similarity of reference metazoan motifs to predicted motifs obtained using a state-of-the-art motif finder that uses promoter sequences [15]. On one hand, PBM profiles cover the spectrum of possible sequences more comprehensively. On the other hand, they include only relatively short motifs. To conclude, no clear winner has yet emerged between PBM technology and traditional motif finding methods in finding PWMs that are closest to reference motifs.

When using binding intensities of one PBM as input and predicting the ranking of probe intensities of another array for the same TF, BE showed best performance. When using PBM binding intensities to predict ranking of promoter intensities in a ChIP-chip experiment for the same TF, SW performed best. We note that there is still only a modest number of TFs with data from both ChIP-chip and PBM; a larger benchmark for *in vivo* prediction, containing also TF binding in metazoans, is needed.

The performance results can be explained by the different goals of the algorithms. RM was designed to optimally rank all probes, so it tries to capture both high-affinity and low-affinity binding information. This explains why it performs less accurately when analyzing the top-binding probes but performs better on very large positive sets (**Figure 4**). The same applies to BE. The inclusion of information from low-intensity binding yields better ranking of low-affinity binding probes, but creates PWMs with lower information content (**Figure 2**). In contrast, AM was designed to identify specific binding motifs; it trains only on the 1000 top-binding 9-mers, and so it only uses information on the specific binding of the protein. Interestingly, SW is best for *in vivo* binding, hinting that longer motifs with a stringent core might be better for this data.

The comparison of the prediction results for *in vitro* and *in vivo* data (**Figure 4**) is striking: The quality of the results is much poorer on *in vivo* data, according to all evaluation criteria (similar results were reported in [20]). This is in spite of the fact that the *in vivo* data consisted of yeast motifs, which are easier to find than mice motifs [5,15]. There can be several explanations of this finding:

1. The length of the probes on the PBM (36bp) is much shorter than the whole yeast promoters targeted by ChIP-chip (an average of 474 bp). As a result, scoring and ranking yeast promoters is harder.
2. Biases caused by the PBM technology lead to systematic distortion in the reconstructed motifs, compared to *in vivo* motifs. If this is the case, revealing and correcting these biases is essential for using the motifs for *in vivo* analysis.
3. The methods tailored specifically for PBMs may overfit this type of data.
4. The complexity of *in vivo* assays distorts the raw binding signals, which look more like the PBM-based motifs in a cleaner *in vitro* environment.

One interesting phenomenon we encountered was secondary motifs: For some PBMs, SW and AM identified a second, completely different motif in addition to the primary one (**Figure S2**). This phenomenon was first reported in [14]. Agius *et al.* suggested that the secondary binding motifs arise as an artefact of the PBM experiment [21]. Zhao and Stormo suggested that secondary motifs are a result of a biased analysis of the PBM data [10], but Morris *et al*. challenge this conclusion [22]. We tested the benefit of using primary and secondary motifs discovered by SW for *in vitro* binding prediction. While there was a significant improvement in performance, it was still worse than BE (data not shown). Jauch et al. recently obtained a crystal structure of the TF Sox4 domain bound to DNA and concluded that two positions in the binding motif are dependent [23]. Such dependency can be manifested by two PWM motifs. Indeed, SW and to some extent AM recover two motifs that reflect this dependence (**Figure S3**). We agree with the conclusion in [20] that more matching PBM and *in vivo* datasets are needed in order to shed more light on this phenomenon.

An interesting insight arises from the comparison of the methods (**Table 4D**). In terms of the Spearman score of probe ranking, all methods are much more similar to each other than to the true binding intensities. This suggests that all methods capture similar information, while missing other pertinent effects (e.g., background or technological biases). On the other hand, predicting the top probes of another method was harder than finding true positive probes (**Table 4D**). Overall, BE had highest pairwise ranking-based scores, concordant with our conclusion that it predicts true binding best (**Table 3**). In terms of distance between PWMs, higher similarities between AM and SW, and between BE and RM, reflect the observation that the former pair produce clear, stringent motifs, while the latter generate more variable, ranking-oriented motifs.

Protein-DNA interactions can occur in a broad range of intensities, and involve both specific and low-affinity (less specific) binding. PBM data enable analysis of the full spectrum of DNA binding affinities of a TF. The binding specificity of a protein can be represented using various models, which differ in expressiveness, compactness, redundancy and interpretability. Our analysis suggests that a PWM models the specific *in vitro* binding quite accurately, obtaining an average AUC of 0.9 on the top probes. The fact that results of all methods tend to deteriorate as the positive sets grow (**Figure 4**), and the success of more complex models in ranking [21] suggest that less specific binding may be better captured by other models. The lower success of all methods in predicting *in vivo* binding questions the transformability of PBM-based results to the *in vivo* domain. Deeper analyses using more data are required on this point.

Our study gauged performance using three criteria: similarity to reference literature motifs, and ability to rank *in vitro* and *in vivo* bindings. The tested methods show a tradeoff between ranking quality and motif similarity. Degenerate motifs are better at *in vitro* binding prediction at the cost of lower information content and similarity to literature motifs. Potential improvement may be achieved by novel methods that strive to optimize both criteria simultaneously.

MATERIALS AND METHODS

Algorithms: We compared four algorithms: Seed-and-Wobble (SW) [4], RankMotif++ (RM) [9], BEEML-PBM (BE) [10] and Amadeus-PBM (AM), a new algorithm presented here (see **Supplementary Methods**). The computational approaches of the algorithms are summarized in

Table 1. Software for BE, RM and SW was downloaded from the authors' websites and run using the default parameters. The full details are in **Supplementary Methods**.

PBM data: We downloaded PBM data from UniPROBE [13]. This database contains, for each TF, paired probe intensity profiles measured on two different arrays. We used the SCI09 dataset, which contains paired profiles of 115 mouse proteins [13,14], and the GR09 dataset, which contains profiles of 89 yeast TFs [19] (**Table 2**).

Reference PWM data: To compare predicted PWMs to experimentally obtained PWMs, we used three databases of reference PWMs: JASPAR [12] and TRANSFAC [16] for mouse motifs and the new yeast motif database ScerTF [17] (**Table 2**). We included in the comparison only reference PWMs that were produced without using PBM data.

ChIP-chip data: We downloaded the ChIP-chip data for yeast TFs from Harbison *et al*. [8]. These data provide large-scale *in vivo* binding for many TFs. Our test used 69 experiments (32 TFs) that had PBM profiles in UniPROBE as well as ChIP-chip measurements.

Comparison and evaluation: We tested the quality of PWMs produced by each method in three ways: by comparison to reference PWMs from the literature (mostly SELEX-based), by their accuracy in predicting *in vitro* binding in PBMs, and by their accuracy in predicting *in vivo* binding as measured by ChIP-chip. In addition, we evaluated how similar the methods are in a pairwise comparison using the same criteria.

To compare a predicted PWM to a reference one, the Euclidean distance between the two PWMs was calculated, as in [15] (for a description of all evaluation criteria see **Supplementary Methods**). The information content of each matrix was also measured in order to evaluate its degeneracy. Each algorithm was trained using the data from both arrays for the same TF. PWMs were also compared using the Tomtom algorithm [24].

For testing the quality of *in vitro* binding prediction, we followed the method of [9]. Since two (paired) binding profiles were available for each TF, a PWM was trained on one profile (the "training array") and used to rank the probes in the other profile (the "test array"). Given a PWM, the probes of the test array were ranked using the sum occupancy score (see **Supplementary Methods**). This ranking was compared to the measured ranking of the probes in the test array according to three criteria: Spearman rank coefficient, sensitivity at 1% false positive rate and area under the ROC curve (AUC) (see **Supplementary Methods** for all definitions). The comparison was done on the probes that showed high binding intensity in the test array (the positive probe set [9]).

To test the quality of *in vivo* binding predictions, we used similar criteria. For each TF, we trained each method using both paired binding profiles (when available) and tested how well the method predicts the ranking of the strongest bound yeast promoters (see **Supplementary Methods**). Predicted and experimental rankings were compared using the same three criteria.

In computing similarity between different methods, we used four criteria. First, we measured the distance between the PWMs inferred by each method. Second, for each method, using the PWM learned on one array, we ranked the set of positive probes in the paired array, and then measured the Spearman rank coefficient between the rankings of each two methods. Third and fourth, we used one method to rank the probes of the paired array, and tested the prediction of the other method using sensitivity at 1% false positive and AUC (see **Supplementary Methods** for computational details).

Statistical significance of the comparison: For each comparison we evaluated its significance using the Wilcoxon rank-sum test [25]. Since the gauged measurements do not distribute normally, we used a non-parametric statistical test.

Supplementary Files

Supplemental Data. Supplementary methods and figures.

Supplemental Table S1. Results of distance to known motifs.

Supplemental Table S2. Results of *in vitro* binding prediction.

Supplemental Table S3. Results of *in vivo* binding prediction.

ACKNOWLEDGEMENT

We are grateful to Quaid Morris for providing us with results of SW and RM runs on the SCI09 data, including trimming uninformative positions. We thank Gary Stormo and Yue Zhao for help with using the BEEML-PBM software and Hershel Safer for comments on the manuscript.

REFERENCES

1. Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2: 28-36.
2. Aparicio O, Geisberg JV, Struhl K (2004) Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. Curr Protoc Cell Biol Chapter 17: Unit 17 17.
3. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316: 1497-1502.
4. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, 3rd, et al. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol 24: 1429-1435.
5. Li N, Tompa M (2006) Analysis of computational approaches for motif discovery. Algorithms Mol Biol 1: 8.
6. Das MK, Dai HK (2007) A survey of DNA motif finding algorithms. BMC Bioinformatics 8 Suppl 7: S21.
7. Sandve GK, Drablos F (2006) A survey of motif discovery methods in an integrated framework. Biol Direct 1: 11.
8. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, et al. (2004) Transcriptional regulatory code of a eukaryotic genome. Nature 431: 99-104.
9. Chen X, Hughes TR, Morris Q (2007) RankMotif++: a motif-search algorithm that accounts for relative ranks of K-mers in binding transcription factors. Bioinformatics 23: i72-79.

10. Zhao Y, Stormo GD (2011) Quantitative analysis demonstrates most transcription factors require only simple models of specificity. Nature Biotechnology 29: 480-483.
11. Tompa M, Li N, Bailey TL, Church GM, De Moor B, et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol 23: 137-144.
12. Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao XB, et al. (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Research 38: D105-D110.
13. Newburger DE, Bulyk ML (2009) UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. Nucleic Acids Res 37: D77-82.
14. Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, et al. (2009) Diversity and complexity in DNA recognition by transcription factors. Science 324: 1720-1723.
15. Linhart C, Halperin Y, Shamir R (2008) Transcription factor and microRNA motif discovery: The Amadeus platform and a compendium of metazoan target sets. Genome Res 18: 1180-1189.
16. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, et al. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31: 374-378.
17. Spivak AT, Stormo GD (2012) ScerTF: a comprehensive database of benchmarked position weight matrices for Saccharomyces species. Nucleic Acids Res 40: D162-168.
18. Tanaka E, Bailey T, Grant CE, Noble WS, Keich U (2011) Improved similarity scores for comparing motifs. Bioinformatics 27: 1603-1609.
19. Zhu C, Byers KJ, McCord RP, Shi Z, Berger MF, et al. (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. Genome Res 19: 556-566.
20. Gordan R, Murphy KF, McCord RP, Zhu C, Vedenko A, et al. (2011) Curated collection of yeast transcription factor DNA binding specificity data reveals novel structural and gene regulatory insights. Genome Biol 12: R125.
21. Agius P, Arvey A, Chang W, Noble WS, Leslie C (2010) High resolution models of transcription factor-DNA affinities improve in vitro and in vivo binding predictions. PLoS Comput Biol 6.
22. Morris Q, Bulyk ML, Hughes TR (2011) Jury remains out on simple models of transcription factor specificity. Nat Biotechnol 29: 483-484.
23. Jauch R, Ng CK, Narasimhan K, Kolatkar PR (2012) The crystal structure of the Sox4 HMG domain-DNA complex suggests a mechanism for positional interdependence in DNA recognition. Biochem J 443: 39-47.
24. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS (2007) Quantifying similarity between motifs. Genome Biol 8: R24.
25. Fay MP, Proschan MA (2010) Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. Stat Surv 4: 1-39.
26. Roider HG, Kanhere A, Manke T, Vingron M (2007) Predicting transcription factor affinities to DNA from a biophysical model. Bioinformatics 23: 134-141.
27. Tanay A (2006) Extensive low-affinity transcriptional interactions in the yeast genome. Genome Res 16: 962-972.
28. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A (1986) Information content of binding sites on nucleotide sequences. J Mol Biol 188: 415-431.
29. Spearman C (2010) The proof and measurement of association between two things. Int J Epidemiol 39: 1137-1150.
30. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognition Letters 27: 861-874.

TABLES AND FIGURES

Table 1: Properties of the tested methods.

Program	Operating principle	Reference
Seed-and-Wobble	Ranks all 8-mers according to Wilcoxon-Mann-Whitney rank-sum score. The top scoring 8-mer is used as a seed, its positions are "wobbled" and its length is extended in order to improve match to the data. http://the_brain.bwh.harvard.edu/PBMAnalysisSuite/index.html	[4]
RankMotif++	Aims to predict the ranking of the probes according to their binding intensity. Maximizes the likelihood of the ranking function, using the three top 7-mers as seeds. http://morrislab.med.utoronto.ca/software.html	[9]
BEEML-PBM	Estimates the position and background biases from the data, then optimizes the parameters of a binding energy model using BEEML algorithm, explicitly taking the biases into account. http://stormo.wustl.edu/beeml/	[10]
Amadeus-PBM	Seeks enriched PWMs in 1000 top ranking 9-mers compared to the background set of all 9-mers, using Amadeus motif finding algorithm. http://acgt.cs.tau.ac.il/amadeus//	Described here

Table 2 Test data and evaluation criteria.

	Data learned on	**Data tested on**	**Test focused on**	**Samples**	**Criteria**
Similarity to known motifs	SCI09 (two arrays)	JASPAR TRANSFAC	Informative positions in the learned PWM	58 80	1. Euclidean distance 2. Tomtom p-value
	GR09 (two arrays)	ScerTF		51	
***In vitro* binding prediction**	SCI09 (one array)	SCI09 (other array)	Top binding probes (4σ-1σ sets, see **Supplementary Methods**)	230 (115 pairs)	1. Spearman rank coefficient 2. True positive at 1% false positive 3. AUC
***In vivo* binding prediction**	GR09 (two arrays)	Harbison *et al*.	Promoters with p-value < 0.0001	69 (out of 89 experiments)	

Table 3. Summary of the comparison. Boldface indicates significantly better performance than the other methods (including equal top performance).

	Similarity to reference motifs	***In vitro* binding prediction**			***In vivo* binding prediction**			**Running time**
	Average Euclidean distance	Spearman rank coefficient	Sensitivity at 1% FP	AUC	Spearman rank coefficient	Sensitivity at 1% FP	AUC	Seconds
AM	**0.178**	0.27	0.342	0.876	**0.152**	0.089	**0.653**	**30**
SW	**0.193**	0.244	0.305	0.866	**0.145**	**0.118**	**0.659**	7200
RM	0.21	0.264	0.295	0.881	**0.158**	0.092	**0.655**	3600
BE	0.227	**0.308**	**0.411**	**0.891**	**0.146**	0.084	**0.665**	900

Table 4. Similarity between methods. (A) For each pair of methods, the Euclidean distance between the PWMs of the two methods is reported. Before the comparison, the column method's PWM is trimmed to eight most informative contiguous positions. (B-D) ranking based comparisons. For each pair of methods, the probe ranking defined according to the column's method is used as reference, and the ranking of the row's method is evaluated using AUC (B) and sensitivity at 1% false positive (C). In (D), for each pair of methods, the 4σ positive sets of the paired PBM are first ranked by each method, and the Spearman rank coefficient of those rankings is computed. In all tables, the average over 230 PBM experiments is reported. Red colour corresponds to greater similarity.

(A)	AM	SW	RM	BE
AM		0.19	0.256	0.249
SW	0.219		0.299	0.245
RM	0.262	0.199		0.183
BE	0.258	0.188	0.179	

(C)	AM	SW	RM	BE
AM		0.877	0.85	0.89
SW	0.876		0.86	0.91
RM	0.843	0.852		0.89
BE	0.877	0.888	0.88	

(B)	AM	SW	RM	BE
AM		0.268	0.192	0.292
SW	0.267		0.232	0.33
RM	0.192	0.228		0.309
BE	0.281	0.325	0.31	

(D)	AM	SW	RM	BE
AM		0.54	0.56	0.65
SW	0.54		0.52	0.63
RM	0.557	0.516		0.65
BE	0.649	0.632	0.65	

Figure 1. Similarity to experimentally established PWMs. For 58 TFs, we compared the motifs produced from their PBM profiles by each method, to the known motif from JASPAR database. Distance was measured using Euclidean distance. Three distance cutoffs were used, and the fraction of recovered motifs with distance below the cutoff is the success rate. BE: BEEML-PBM, RM: RankMotif++, SW: Seed-and-Wobble, AM: Amadeus-PBM, JR: JASPAR.

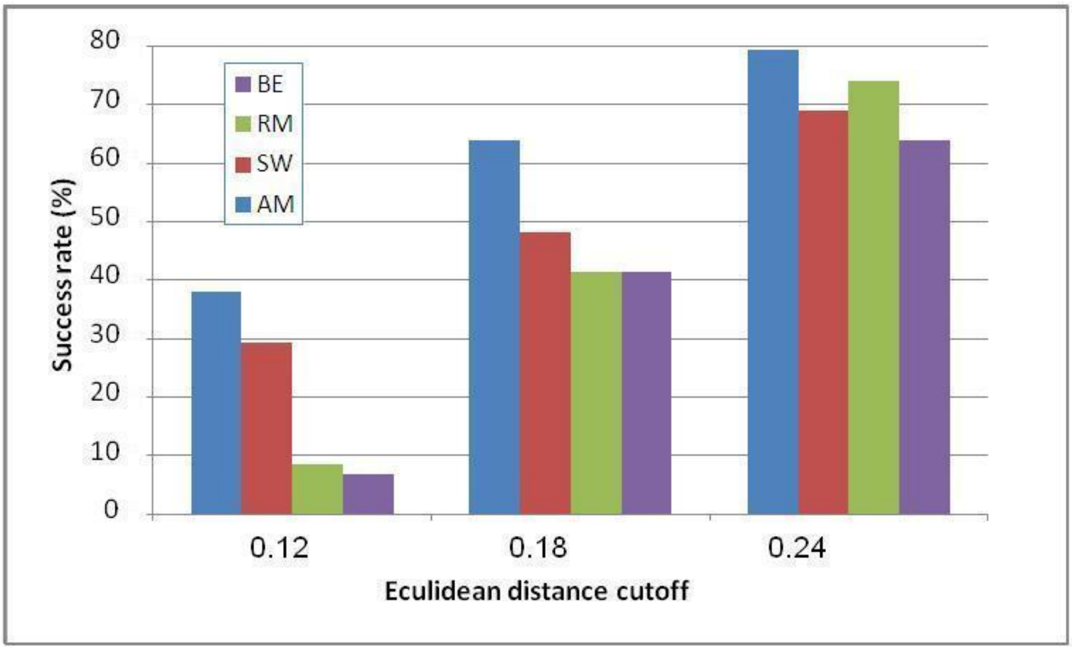


Figure 2. Examples of generated motifs. The figure shows examples of the motifs produced by each method and the corresponding JASPAR motif. For three proteins, the PWM logos produced by each method and the experimentally and independently established motif in the JASPAR database are shown. AM was trained on motif length 8, while for BE, RM and SW only the most informative contiguous positions were kept. We chose TFs whose motifs had information content most similar to the averages of the different methods.

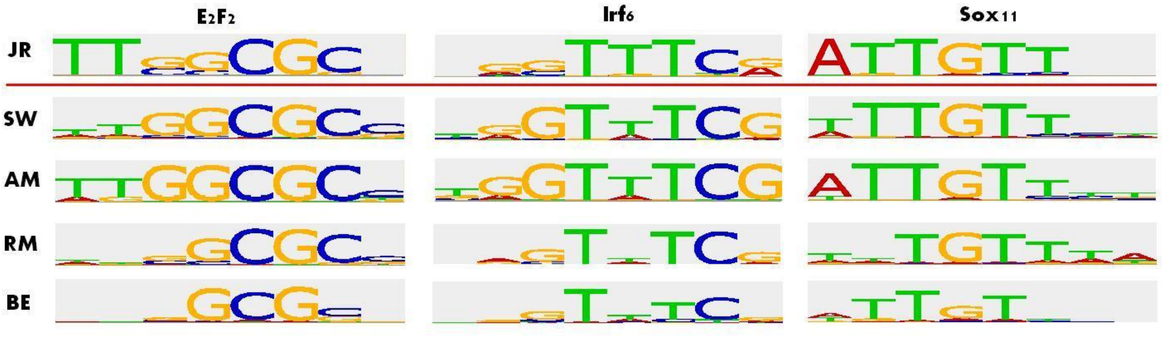


Figure 3. Success rates in probe ranking of a paired PBM. For each TF and method, the PWM was learned using one array and used to infer probe intensity ranking in its paired array. Ranking was gauged on a set of top positive probes (4σ set) according to three measures: Spearman rank coefficient, sensitivity at 1% false positive and AUC (see **Supplementary Methods** for all mathematical terms). For each quality measure, three distance cutoffs were used, and the fraction of TFs with score equal or better to the cutoff is the success rate. The results show the success rate over 230 samples (115 paired arrays).

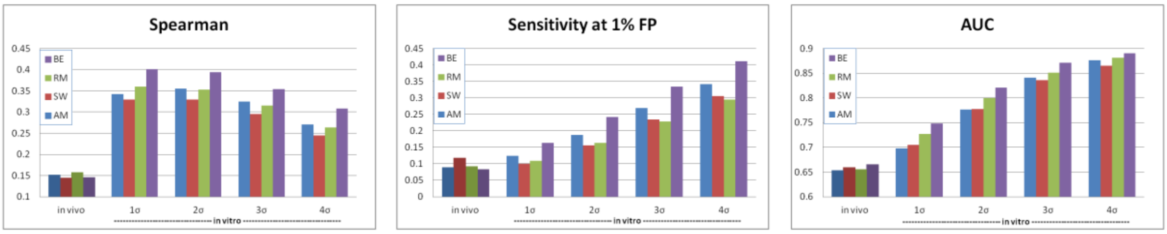


Figure 4. Quality of binding prediction for *in vivo* and *in vitro* data of different sizes. For each of the four algorithms, the quality of the motifs inferred from PBMs in ranking the top binding probes as measured *in vivo* (by ChIP-chip experiments) and *in vitro* (by PBMs) was evaluated. The *in vivo* test included 69 yeast ChIP-chip experiments data (with an average of 61 promoters per experiment). The *in vitro* test included 230 mouse PBMs covering 115 TFs, and used several definitions for the sets of top binding promoter sequences (4σ to 1σ, with averages of 912, 1580, 3215 and 8224 top probes, respectively, see text). Ranking quality was measured by the Spearman rank coefficient, the sensitivity at 1% false positive (FP) and the area under the ROC curve (AUC) (see **Supplementary Methods**).The average ranking quality is reported in each case.

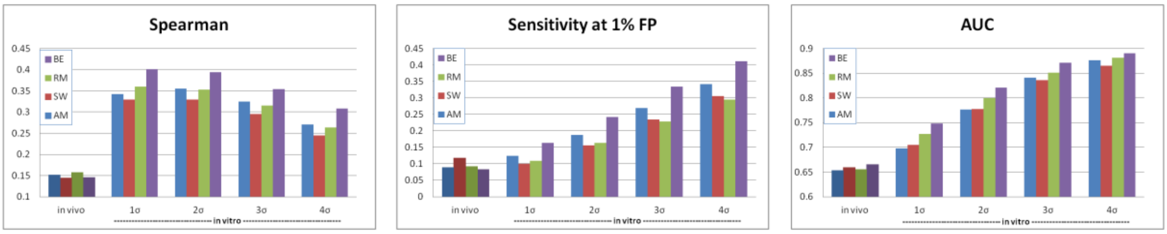


Figure S1. Similarity to experimentally established PWMs. (A) TRANSFAC motifs. For 80 proteins available in TRANSFAC we compared the motifs produced from their PBM data by each of the tested methods to the motif available in TRANSFAC. Distance was measured using Euclidean distance. Three distance cutoffs were used, 0.12, 0.18 and 0.24, and the fraction of recovered motifs with distance below the cutoff is the success rate. (B): ScerTF motifs. The same tests on 51 motifs from the ScerTF database. AM: Amadeus-PBM; SW: Seed&Wobble; RM: Rankmotif++; BE: BEEML-PBM.

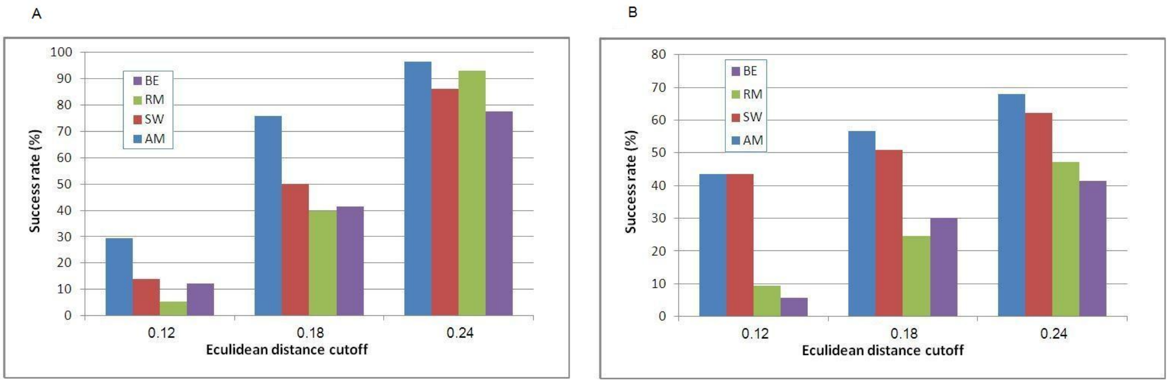


Figure S2. Shadow motifs. Examples of the primary and secondary motifs found by Amadeus for Pou2f3 (A) and Sox1 (B). p-values for the motif enrichment (hypergeometric score) are indicated above each motif. Note that even the second ranked motifs obtain extremely high significance.

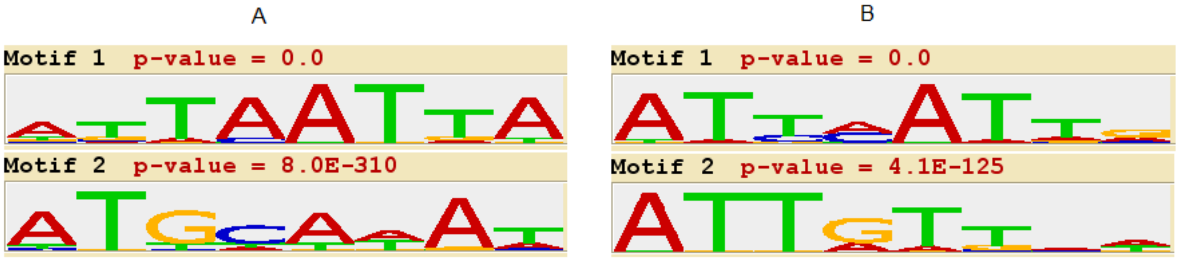
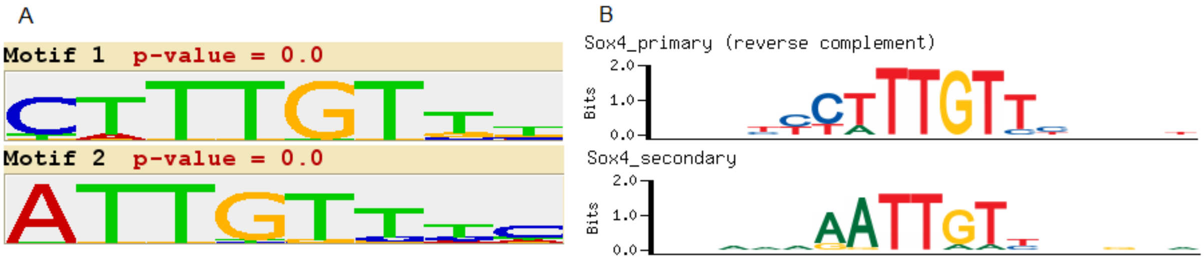


Figure S3. Sox4 primary and secondary motifs as found by Seed-and-Wobble (SW) and Amadeus-PBM (AM). Jauch et al. reported two motifs: CTTTGTT and AATTGTT (23). (A) The two top motifs recovered by AM. The first motif of Jauch et al. was recovered correctly; the second was partially recovered. (B) The two top motifs recovered by SW. Both motifs from Jauch et al. were inferred correctly. Logos taken from UniPROBE database (13).



Supporting Information for

Assessment of algorithms for inferring positional weight matrix motifs of transcription factor binding sites using protein binding microarray data

by Orenstein, Linhart and Shamir

Supplementary Methods

Algorithms and parameters

BE, RM and SW were downloaded from the authors' websites and applied using the recommended parameters. The parameters for SW were k-mer length 8, seed pattern file patterns8of12, total pattern file patterns_4x44k_all_8mer. The parameters for RM were widths to try 7 to 13, no log transformation, optimize the scaling factor w, 400 negative sequences. The PWM with the best likelihood score was taken as the result. BE was run according to the R script provided on the authors' website. Its seed PWM was the SW result trained on the same array. Amadeus was run on the top 1000 9-mers as the target set, motif width 8 and all other parameters at default values. In particular, the background set included all 9-mers. For computing the Spearman rank coefficient, sensitivity at 1% false positive and AUC, the algorithms were run on one array and tested on the other. In the comparisons to known motifs and *in vivo* data, the algorithms were trained on both arrays together.

Amadeus-PBM algorithm

We devised a simple scheme for detecting TFBS motifs in PBM data. The method is generic in that it can utilize any motif finding algorithm and any ranking score. Enrichment-based motif finding algorithms receive as input a target set of sequences that are expected to be enriched with the motif compared to other (background) sequences. Our approach to utilize such algorithm is quite simple:

1. Rank all *k*-mers according to some score that reflects their binding intensity.
2. Give the top *N* ranking *k*-mers as the target set to the motif finding algorithm, using all *k*-mers as the background set.

The rationale for using k-mers is that TFBSs have typically short motifs that will be reflected in overrepresented k-mers among high binding intensities. A PBM contains each 10-mer once, so that each non-palindromic 10-mer will appear twice, once on each strand. By choosing $k < 10$, each k-mer appears several times in the probe set, and the mean (or median) of the intensities of the probes containing these multiple occurrences can be used to rank the k-mers, thereby reducing noisy measurements and possible biases (due to, e.g., position in the probe sequence, flanking sequences and strand). We found the average binding intensity the most suitable. Note that the method works on the original binding intensities as reported in the PBM data. We found that using k=9 improved the accuracy of the results over k=8 when looking for a motif of length 8, since the set of top-ranking 9-mers may contain several shifts of the same 8-long motif (data not shown). Taking N=1000 proved to be a good compromise between adding more noise and leaving out too many 9-mers with true positive binding sites. Our motif finding algorithm of choice was Amadeus (15). We call the resulting method Amadeus-PBM.

Mathematical criteria for evaluation of motif quality

Position weight matrix (PWM): The TFBS model used by all tested algorithms is a PWM: It is a *4×k* matrix *Θ*, where *$\Theta_i(x)$* is the probability of base *x* in a position *i* of the model. *k* varies with the motif. The score of k-mer $w_1...w_k$ is $\prod_{i=1}^{k}\Theta_i[w_i]$.

Distance score: To measure the distance between an inferred PWM to a reference PWM, we used *Euclidean distance* [8]. The Euclidean distance of probability vectors *(v_1, v_2, v_3, v_4)* and *(u_1, u_2, u_3, u_4)* where $\sum_i v_i = 1$ and $\sum_i u_i = 1$ is:

$$d(u,v)=\sqrt{\sum_{i=1}^{4}(u_i-v_i)^2}$$

If p_1 and p_2 are two aligned PWMs of length *k*, where p_{1i} is the *i*-th column of p_1, the Euclidean distance is

$$e(p_1,p_2)=\frac{1}{\sqrt{2}k}\sum_{i=1}^{k}d(p_{1i},p_{2i})$$

Hence, *$0\leq e\leq 1$* with smaller values indicating higher similarity.

Given two PWMs, all possible (gap-free) alignments between them in both orientations with an overlap≥5 are tested, and the smallest obtained score is defined as the distance between the PWMs.

Occupancy score: Evaluating the chance that a PWM *Θ* binds to a probe or a promoter sequence *s* is done by summing the probabilities of all possible alignments of *Θ* to *s*. Formally, the *sum occupancy score* [26] for sequence *s* and PWM *Θ* is defined as

$$f(s,\Theta)=\sum_{t=0}^{|s|-k}\prod_{i=1}^{k}\Theta_i[s_{t+i}]$$

Taking the sum was reported to give the better results than taking the maximum [27].

Positive probes: To evaluate how well a motif predicts the binding of probes in a PBM, one has to focus on the strongest, specific binding, as the lower binding intensities may be noisy and non-specific. Given the binding intensities of a protein to all probes in a PBM, Chen *et al.* [9] defined the *positive probe set* as those probes whose normalized binding intensity is greater than the median by at least 4 * (MAD / 0.6745), where MAD is the median absolute deviation (MAD = 0.6745 for the normal distribution *N(0,1)*). We denote σ = (MAD / 0.6745). In some tests we also used larger positive probe sets by setting the threshold to 3σ, 2σ, and 1σ above the median.

Positive promoters: To evaluate how well a motif predicts the binding of a TF in a ChIP experiment, one has to identify the specific bindings. Haribson *et al.* [8] defined the *positive promoter set* as those promoters whose reported p-value for binding to the TF is smaller than 0.001.

Information content: We used the entropy to measure the information content of each PWM [28]. The bit information of vector *(v_1, v_2, v_3, v_4)* (where $\sum_i v_i = 1$) is defined as $2+\sum_i v_i \log(v_i)$. The information content of a PWM is the average bit information of its columns.

Ranking criteria: To evaluate the probe (or sequence) ranking of an algorithm, we used the same three criteria as in [9]. In all cases, we have *n* probes ranked $x_1\leq x_2\leq\ldots\leq x_n$ according to some algorithm, while the true ranking according to binding intensities of probe *i* is *y_i*. The **Spearman rank coefficient** [29] compares how similar the two rankings of the positive probe set are, using the formula:

$$1-\frac{6\sum_{i=1}^{n}(x_i-y_i)^2}{n(n^2-1)}$$

Suppose there are P positive and N negative probes. Denote z_i=1 if probe *i* is in the positive set and z_i=0 otherwise. If the algorithm assigns the top *t* probes as positive, there are $TP(t)=\sum_{j=1}^{t} z_j$ *true positive* and $FP(t)=t-TP(t)=\sum_{j=1}^{t}(1-z_j)$ *false negative* samples. Of the remaining probes, which are declared negative, $FN(t)=P-TP(t)$ are *false negative* and $TN(t)=N-FP(t)$ are *true negative*. The *sensitivity* is defined as $TPR(t)=TP(t)/P$ and the *false positive rate* is $FP(t)/N$. Let *k* be the maximum number of top ranking probes that attain 1% false positive rate. Then the **sensitivity at 1% false positive** is defined as $TPR(k)$. The *receiver operating characteristic* (ROC) curve is defined as the function plotting the sensitivity against the false positive rate as *t* increases. The area under the ROC curve or **AUC** is used as the third criterion [30].

Supplementary Results

Comparing predicted motifs to TRANSFAC and ScerTF

We compared the motifs predicted by each method to PWMs reported in the TRANSFAC and ScerTF databases, in the same fashion as the comparison to JASPAR motifs. There were 80 TRANSFAC PWMs and 51 ScerTF PWMs corresponding to TFs from the SCI09 and GR09 studies, respectively, which did not originate from PBM data. Dissimilarity was measured using Euclidean distance. The results were clearly in favor of AM, followed by SW, RM and BE in this order (**Figure S1**). Full results are available in **Table S1**.

Comparing the significance of the similarity to literature motifs

As an additional quality measure, we scored the motif similarity using the recently developed Tomtom algorithm (24), which calculates the significance of the similarity. Given a query motif and a motif database, Tomtom outputs a p-value for the similarity of the query to different motifs in the database. AM showed the highest significance levels. For example, the number of PWMs with p-value below 0.01 threshold detected in the JASPAR database for AM, SW, RM and BE were 46, 40, 43 and 38, respectively. Results are available in **Table S1**.