MARS: Motif Assessment and Ranking Suite for transcription factor binding motifs

Caleb Kipkurui Kibet* and Philip Machanick

Department of Computer Science and Research Unit in Bioinformatics (RUBi), Rhodes University, Grahamstown, 6140, South Africa

*** calebkibet88@gmail.com**

Abstract

We describe MARS (Motif Assessment and Ranking Suite), a web-based suite of tools used to evaluate and rank PWM-based motifs. The increased number of learned motif models that are spread across databases and in different PWM formats, leading to a choice dilemma among the users, is our motivation. This increase has been driven by the difficulty of modelling transcription factor binding sites and the advance in high-throughput sequencing technologies at a continually reducing cost. Therefore, several experimental techniques have been developed resulting in diverse motif-finding algorithms and databases. We collate a wide variety of available motifs into a benchmark database, including the corresponding experimental ChIP-seq and PBM data obtained from ENCODE and UniPROBE databases, respectively. The implemented tools include: a data-independent consistency-based motif assessment and ranking (CB-MAR), which is based on the idea that 'correct motifs' are more similar to each other while incorrect motifs will differ from each other; and a scoring and classification-based algorithms, which rank binding models by their ability to discriminate sequences known to contain binding sites from those without. The CB-MAR and scoring techniques have a 0.86 and 0.73 median rank correlation using ChIP-seq and PBM respectively. Best motifs selected by CB-MAR achieve a mean AUC of 0.75, comparable to those ranked by held out data at 0.76 – this is based on ChIP-seq motif discovery using five algorithms on 110 transcription factors. We have demonstrated the benefit of this web server in motif choice and ranking, as well as in motif discovery. It can be accessed at `www.bioinf.ict.ru.ac.za`.

Introduction

We introduce MARS (Motif Assessment and Ranking Suite), a web server hosting a suite of tools for motif evaluation and ranking. It provides a service

that is necessitated by the advance in high-throughput sequencing technologies at a continually reducing cost that has seen a large amount of – often noisy – data being generated by various studies [27]. To make sense of these datasets, several tools and algorithms have been developed, differing in data cleaning and statistical algorithms involved. The wide variety and the large number of computational tools being developed makes it hard for a non-specialist with limited computational skills to choose the best tools for use in their research [1]. Additionally, to improve on currently available tools, algorithm developers need well thought out and representative benchmark data (gold standard) and evaluation statistics. This problem has been tackled by independent evaluation studies [24, 25, 35, 37] focused on various niches of research and data, producing incomparable results. This prompted a question of "who watches the watchmen" (evaluation benchmarks) by Iantorno *et al.*, [14] who also proposed that a proper benchmark should follow a set of pre-determined criteria to ensure that the evaluations are biologically relevant. Aniba *et al.*. [1] provide a set of criteria for a good benchmark that we adapt to the context of motif assessment as follows:

- **Relevant** – scoring methods should provide biologically meaningful evaluations
- **Solvable** – scoring methods should not be trivial but must be possible to use with reasonable effort
- **Scalable** – the benchmark should be expandable to cover new techniques and algorithms as they develop
- **Accessible** – data and statistical tools should be easy to source and use to evaluate other algorithms or protocols
- **Independent** – methods should not be tailored or biased to a particular algorithm or biased towards certain experimental techniques
- **Evolvable** – the benchmark should change as new data are made available, as well as to reflect the current problems and challenges in the field

The evaluation problem has been widely investigated in multiple sequence alignment, 3D structure prediction, protein function and gene expression analysis [1], mostly following the criteria above. However, evaluation remains an active challenge in gene regulatory research, especially in predicting TF binding sites and the accuracy of prediction models [39]. This difficulty is directly linked to the motif discovery problem, which has been attributed to the degeneracy of TF binding and the presence of multiple potential binding sites in the genome [13, 16]. The difficulty of motif discovery has in turn driven the growth in experimental techniques developed to improve the affinity and specificity of TF binding site prediction models; techniques to identify binding sites or binding affinity include Chromatin Immunoprecipitation followed by parallel sequencing (ChIP-seq) [15] or exonuclease cleavage (ChIP-exo) [28], protein binding microarray (PBM) [2], Assay for Transposase-Accessible Chromatin

with high-throughput sequencing (ATAC-seq) [7], DNase I digestion and high-throughput sequencing (DNase-seq) [31] and many others. Consequently, the number of algorithms and hence the binding models in databases continues to increase. Two areas are in need of evaluation: the algorithms used in motif discovery and the models deposited in the various motif databases. Although interlinked, in that ranking a model can be an indirect evaluation of an algorithm used to generate it, most of the evaluation attempts so far have been focused on the algorithms. This is a challenging task given that new tools are published regularly with varied implementations, scoring functions and even the data used for motif discovery. Therefore, establishing a widely useful model or motif algorithm evaluation platform is a moving target.

Nonetheless, there have been some attempts to develop tools and techniques to evaluate motif discovery algorithms, which can be categorized into assess-by-binding site prediction, motif comparison or by sequence scoring and classification [18]. We relate a selection of known approaches to the benchmark criteria we outline above.

An assess-by-binding site prediction approach evaluates algorithms by their ability to identify known or inserted binding sites in a sequence. It is widely investigated; a number of stand-alone motif assessment tools [26] and web servers [30, 33] have been developed. However, these tools neither *evolved* nor *scaled* with advances in motif discovery algorithms, reducing their *relevance*, and thus failing to meet major requirements for an evaluation benchmark. Assess-by-scoring and classification tests binding models by their ability to discriminate sequences known to contain binding sites from those without. UniPROBE is the most comprehensive collection of PBM-derived motifs [22] and we are aware of one web server that uses such data in motif evaluation [36] (it has neither *evolved*, *scaled* not is it easily *accessible*) while, for ChIP-seq data, Swiss Bioinformatics hosts a simple web server (PWMTools (`http://ccg.vital-it.ch/pwmtools/pwmeval.php`) that is limited to testing single motifs against ENCODE data using sum occupancy scoring; it does not allow for comparative motif or sequence data testing (not *relevant* or *accessible*) and the site is not published. On the other hand, assess-by-motif-comparison has generally been used to determine if the discovered motifs are similar to those in 'reference databases' using motif comparison algorithms. An algorithm is considered successful if it can predict a motif similar to those in the database. However, this assumes the accuracy of previous predictions (not *relevant* or *scalable*), a weakness we address in this study.

The spread of motif models across DBs and in different PWM formats makes it difficult to create a benchmark that ranks multiple motifs for a given TF, and this problem is compounded by the growth in available data [18]. There is a lack of an easily *accessible* and *independent* motif evaluation platform that can allow users to rank PWM models for a given TF. To fill this gap, we introduce a web server that hosts a suite of motif assessment tools used to evaluate and rank motifs. For wider applicability, we collect ChIP-seq and PBM data generated from different labs and use an average score to represent a given motif, with the assumption that this would capture the most general binding behaviour. We

also apply a wide variety of scoring functions and statistics to reduce technique bias. In addition, we introduce a novel Consistency-Based Motif Assessment and Ranking (CB-MAR) approach that can be considered to be *data-independent*, hence less biased compared to scoring-based techniques.

Materials and Methods

Benchmark data

We downloaded all ChIP-seq peaks uniformly processed by Analysis Working Group (AWG) from ENCODE [32], PBM from UniPROBE [5] database and PMW motifs from various databases and publications, prepared as previously described [18] and stored in a MySQL database (Table 1). Alternative TF names (from GeneCards [29]) link various alternative TF names to a TF class ID derived from the TFClass classification [38] to find all motifs for a TF irrespective of naming inconsistency. Unless otherwise specified, all sequence data used are derived form the human genome (hg19), except for the PBM data where we also use mouse data.

Table 1. Summary of Benchmark Data in the Database: 'Used in analysis' represent data using in our comparative tests. For motif discovery with GimmeMotifs we use 110 (in brackets).

Data Type	TFs	motifs	motifs per TF	Used in Analysis
ChIP-seq	161	691	4.3	83 (110)
PBM	285	455	1.6	60
PWM-Motifs	1352	6050	4.5	3686

Algorithms Overview

We have previously described the implementation of assessing by scoring and enrichment algorithms [18]. In summary, for each TF, the motifs in PWM format are used to score sequences partitioned into positive (test) and the negative (background) using one of the implemented scoring functions. Finally, the ability to classify the two sets is evaluated by area under receiver operator characteristic curve (AUC) or the mean normalized conditional probability (MNCP) statistics (Table 2). See our previous paper [18] for more details on the scoring functions and statistics used.

Consistency-Based Motif Assessment and Ranking (CB-MAR)

CB-MAR is based on the idea that 'correct motifs' are more similar to each other while incorrect motifs will differ from each other. The logic for this view is that differing methods are unlikely to reproduce each others' errors. This idea is used in evaluating sequence alignments: correct ones are assumed to compare with each other in a consistent manner, while incorrect ones will differ from each other in various ways, generating inconsistent alignments [14, 20].

Table 2. Implemented scoring functions: The table provides a list of scoring functions used and the recommended statistics.

Scoring functions	Preferred stat.	Recommended?	Reference
Energy	AUC or MNCP	High (Default)	[40]
GOMMER	MNCP	High	[8]
Sum Occupancy	MNCP	Average	[23]
Max Occupancy	MNCP	Average	[41]
Max Log-odds		Not Recommended	[41]
Sum Log-odds		Not Recommended	[3]

We implement this approach using Tomtom [12] and FISim [11] motif comparison algorithms. For a given TF, we calculate a similarity score between all motif pairs and finally an average motif similarity score, which we use as a measure of motif quality. For best results, the benchmark motif set should be: (a) generated from a variety of data and motif finding algorithms – with (b) identical motifs eliminated (especially in a small set) – and (c) be large enough to capture variation in binding behaviours of the TF. The optimum number depends on the TF: one with uniform behaviour can be characterised with a smaller set of motifs than one with variable binding affinity, for example.

In more detail, CB-MAR is implemented as follows. Given a TF with a collection of motifs M of size n and using Tomtom's Euclidean distance (ED) for motif comparison, we define Pairwise Similarity Score (PSS) based on Tomtom P-value P_{M_i,M_j}. The PSS for motif M_i and M_j is computed as:

$$PSS(M_i, M_j) = -log(P_{M_i,M_j}), \tag{1}$$

and then normalized by the maximum score of all PSS scores of M_i. The Average Similarity Score (ASS), which we use as the measure of quality and rank, of motif M_i, is then computed as:

$$ASS(M_i) = \frac{\sum_j^n PSS_{(M_i,M_j)}}{n}, \tag{2}$$

MARS Web server Implementation

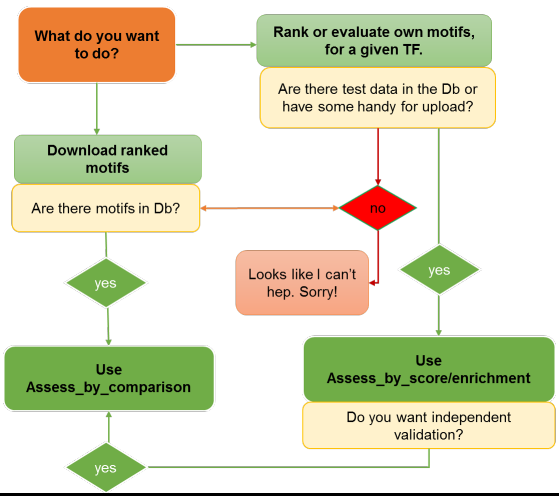


Figure 1. Decision flow diagram. Guides the user on the appropriate tools to use in MARS:

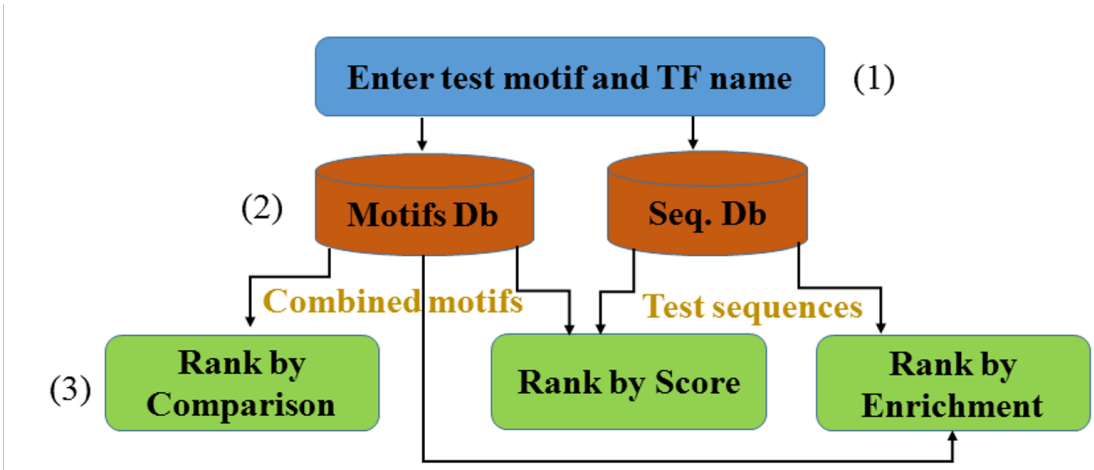
The MARS web server is implemented in Django, a Python web framework, and hosted on an Apache web server while the PWM motifs and sequence benchmark data are stored in a MySQL database. MARS is designed to allow the users to either retrieve ranked motifs for a given TF or rank their own, as long as the required test data is available or uploaded. A guided

search function for the available motif and benchmark data assists the users when choosing the tools to use, based on the decision flow diagram in Figure 1. The MARS documentation also provides a detailed guide on the accepted data formats and best practices when using the various tools in the web-server (`www.bioinf.ict.ru.ac.za/documentation`). For any analysis, the TF name is the only required input, used to retrieve the available motifs from the database and benchmark data in the case of assess-by-score and enrichment methods (Figure 2). Where the data is not available, the user is prompted to upload. MARS currently accepts motifs in MEME format and test ChIP-seq data in BED format.

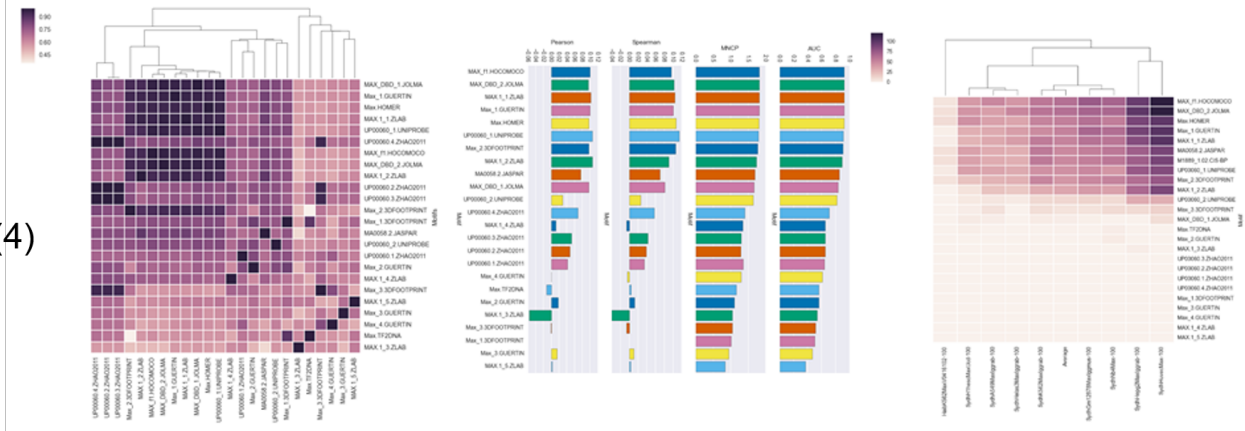
Evaluation of MARS tools

Comparison of the assessment approaches: How well tools implementing different algorithms and data reproduce each other can act as a crude evaluation. For our evaluation, we select a total of 127 TFs that have a TFClass ID, have more than 10 motifs, and have benchmark data sourced from either PBM (60 TFs) or ChIP-seq (83 TFs) (see Table 1) to rank all the available motifs for each TF using the different tools available. For simplicity of analysis and comparison, we use energy scoring function and AUC statistics throughout these evaluations (Table 2) – a combination we found [18] to produce consistent rankings and is least biased by motif length and information content (IC).

Motif assessment in *ab initio* motif discovery: To validate CB-MAR, we apply it to choose the best motifs in *ab initio* discovery, a task commonly accomplished using held out data. We take advantage of an ensemble motif finding tool, GimmeMotifs [34], which performs *ab initio* motif discovery from ChIP-seq data using nine algorithms. A total of 110 TFs, which had ENCODE ChIP-seq data and a corresponding TF-class ID, were chosen for the analysis. From all the available data from different cell lines for a given TF, we extract the top 500 peaks widened around the peak centre to 100 bp, merged and shuffled to avoid sampling bias when partitioned. We then perform *ab initio* motif discovery on 50% of the data and the rest is held out for validation. We randomly sampled 5000 peak sequences for Ctcf as it had a large data set to reduce computational costs. After motif discovery, we use our CB-MAR (using Tomtom) approach in combination with motif clustering (using *gimme cluster* from GimmeMotifs at 95% similarity) to rank and narrow down the motif predictions to the best three non-redundant motifs. Next, we use the validation data to evaluate the best motifs identified by CB-MAR and GimmeMotifs using the *gimme roc* command. Internally, GimmeMotifs uses 20% of the input sequences for discovery and the rest for evaluation and ranking.



(4)



(5)

Click on the table headers to sort

Motif_name	Motif_IC	Average_IC	Motif_length	Motif_score ▲	Motif_logo
Max.HOMER	11.292923	1.026629	11	9.696135	Ceqlogo 10.11.15 17:56
MAX.1_1.ZLAB	13.242696	1.018669	13	9.285802	Ceqlogo 10.11.15 17:56

Figure 2. Motif assessment flow diagram: (1) User enters a TF name and/or uploads motifs in MEME format (to evaluate own data). (2) Motifs and test sequences linked to the TF are extracted from the database. (3) Motifs can be ranked, by comparison, used to score test sequences (rank by score) or its enrichment determined using CentriMo (rank by enrichment). (4) The results are visualized interactively, (5) with additional information like motif length, Information content, and logo. The clustergram offers additional details on the motif or test data clustering. In the end, the user can download ranked motifs in MEME format, as well as raw data for further analysis.

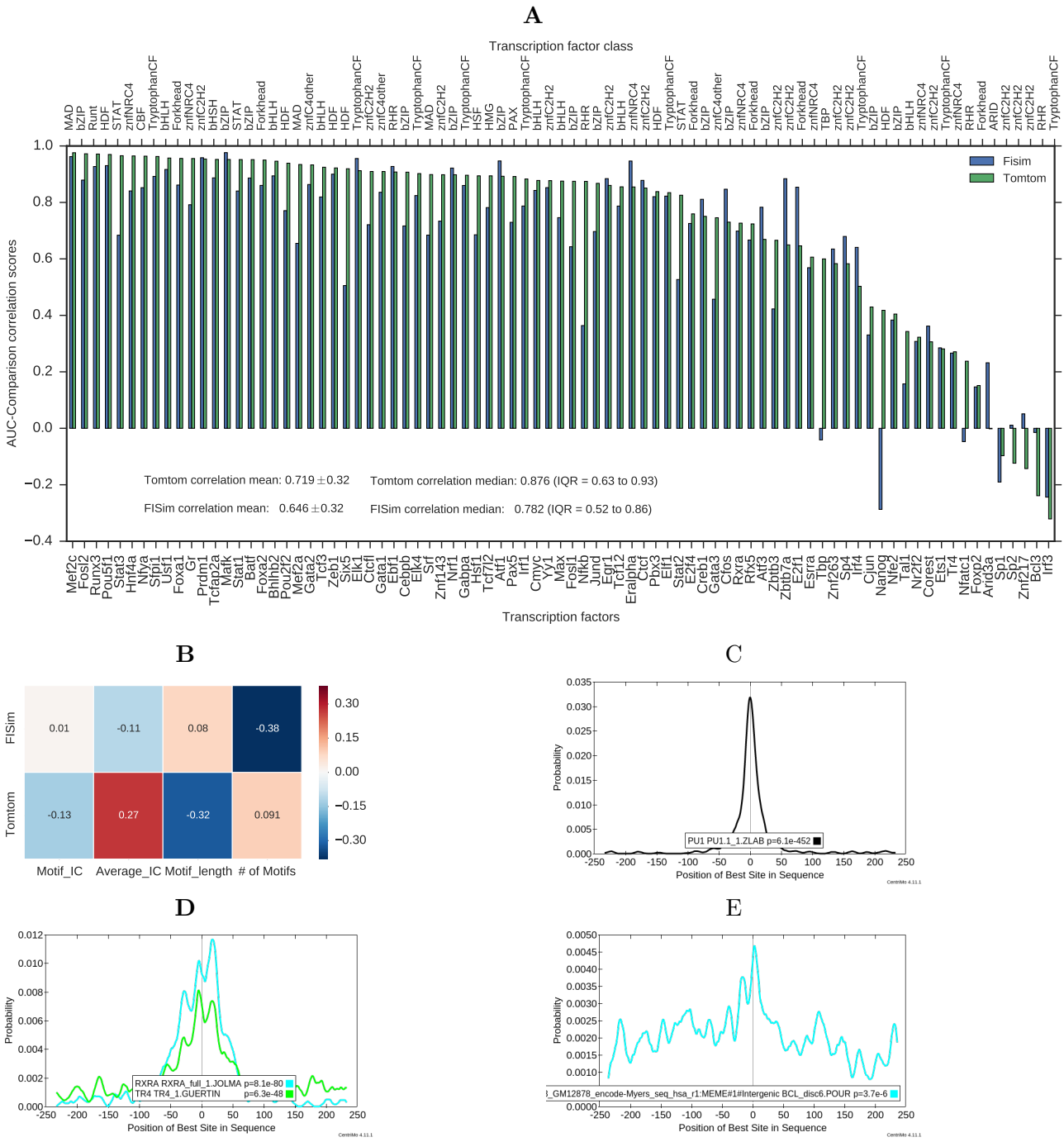


Figure 3. Correlating consistency-based assessment (Tomtom and FISim) with Energy AUC ranking in ChIP-seq data: The bar graph **(A)** shows how rankings based on Energy scoring correlate with consistency based techniques. The mean±STD and median with interquartile range statistics are annotated on the graph. In **(B)**, we show the effect of motif information content (total and averaged by length), the number of motifs (size) and length on motif ranking. The CentriMo plots predict the possible direct binding behaviours (based on the sharp, centred peaks) of **(C)** Pu1 motifs in ChIP-seq peaks, and indirect or cooperative binding of **(D)** Tr4 and **(E)** Bcl3 motifs. The motif names and the *p*-value of central enrichment of the ChIPed motifs is provided in the figure legends. For Tr4, the other centrally enriched motif (Rxra) could bind cooperatively with it.

Results

Benchmarking and evaluation of algorithms

For the assess-by-score approach, we have previously performed a thorough comparison and testing to determine the effect of the scoring functions and motif characteristics (length and IC) on score and rank [18]. In summary, we found that motif ranking is influenced by the scoring function used in a TF-specific manner with energy scoring producing the most 'biologically relevant' rankings and that motif IC is not a reliable predictor of motif quality (Table 2). Since there is no 'ground truth' by which we can evaluate these motif assessment approaches, we rely on how well rankings from the different tools agree and how well our benchmarks and evaluations meet the requirements listed in the introduction.

For assess-by-comparison, we determine whether CB-MAR is influenced by motif length or information content, which could create a bias, and see if that could explain the difference in performance between Tomtom and FISim. Tomtom scores have a positive correlation (R=0.24, favouring higher IC motifs) with average IC normalized over motif length while FISim scores have a negative correlation (-0.11, penalize higher IC motifs). FISim is not influenced by length (R=0.078) while Tomtom penalizes longer motifs (R=-0.25) since average IC is lower for longer motifs with low IC flanks. Surprisingly, the number of motifs for the TF seems to negatively affect motif scores in FISim (R=-0.38) – due to higher IC as the number of motifs is increased – but has no effect for Tomtom (R=0.011), possibly explaining their difference in performance (Figure 3B).

Tomtom comparison produces 'biologically relevant' rankings

For consistency-based motif ranking (CB-MAR), we decide on the best motif comparison algorithms that generate biologically relevant rankings – as defined by how well the motif ranks reproduce those based on *in vivo* data (ChIP-seq) – by correlating with ranks based on energy scoring. From figure 3A, we observe that the scores and ranks based on Tomtom (median=0.88; Interquartile range, IQR=0.63-0.93) can better reproduce AUC scores based on energy scoring compared with FISim (median=0.78; IQR=0.52-0.86). We use the median to summarise the performance of the two techniques since the correlation scores were skewed (Figure 4). The level of correlation between CB-MAR and energy AUC rankings seem to also predict the binding behaviour of the TFs. For Tomtom, we find highly

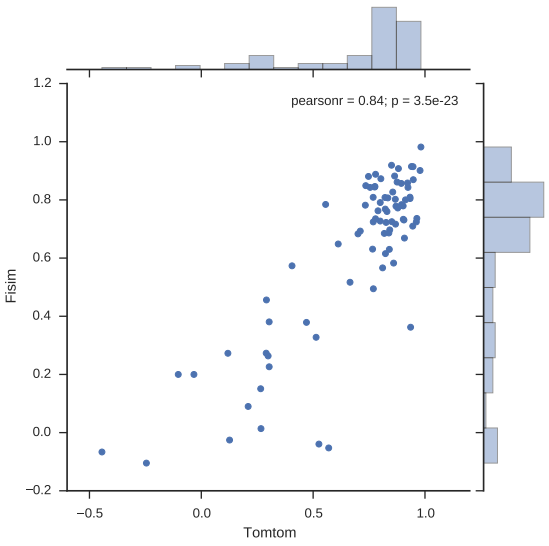


Figure 4. Joint scatter plot and histogram shows the skewed distribution of Spearman rank correlation scores of Tomtom and FISim with those based on Energy scoring

correlated motifs also have centrally enriched peaks in CentriMo (Figure 3B and C) while less or negatively correlated TFs have broad peaks (Figure 3D and E), a known predictor of indirect or cooperative binding [4]. The most common poorly correlated TF family, znfC2H2, are also known to bind in a sequence independent manner [16].

For in vitro data (PBM), however, we do not find a clear performance difference between the median average AUC scores in Tomtom (0.70) and Fisim (0.72), but observe a higher mean in FISim (0.56) compared with Tomtom (0.52), hinting that FISim could better model *in vitro* while Tomtom models *in vivo* binding better. We also note that the TFs with low correlation between scores are known to bind indirectly or cooperatively. Specifically, the TFs from high mobility group (HMG) which have a negative correlation, are known to bind both directly and cooperatively, but they may have a different binding behaviour *in vivo* and *in vitro*. Finally, for TFs with data from both ChIP-seq and PBM (21 TFs), we compared how the Energy scores and motif rankings in the two data types correlate. We observe a similar trend, where the correlation scores reflect the TF binding behaviour (Figure 5). Zbtb3 has a negative correlation of over -0.7, a possible indicator of a difference between *in vivo* and *in vitro* binding behaviour.

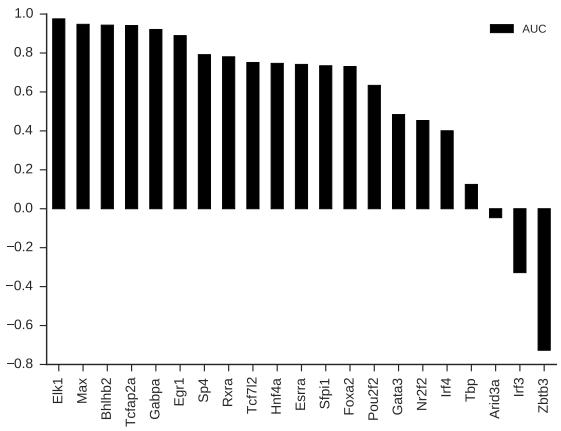


Figure 5. Zbtb3 binds differently *in vivo* and *in vitro*: Motif ranks are in ChIP-seq and PBM benchmark data are mostly in agreement except for a few TFs.

Assess-by-score-Energy reproduces *gimme roc* rankings

GimeMotifs [34], an ensemble motif discovery pipeline for ChIP-seq data, also includes *gimme roc* for motif quality analysis and ranking. We use this to benchmark our approach and found that *gimme roc* produces motif rankings significantly correlated with the Energy scoring ranks (R=0.99 Pearson, p=1.9×10^{-105}) (Figure 6) and **(R=0.995 Spearman's, correlation p=1.7×10^{-108})**. Also, there is no significant difference between the two sets of scores (**p=0.825, Wilcoxon rank-sum test)**,

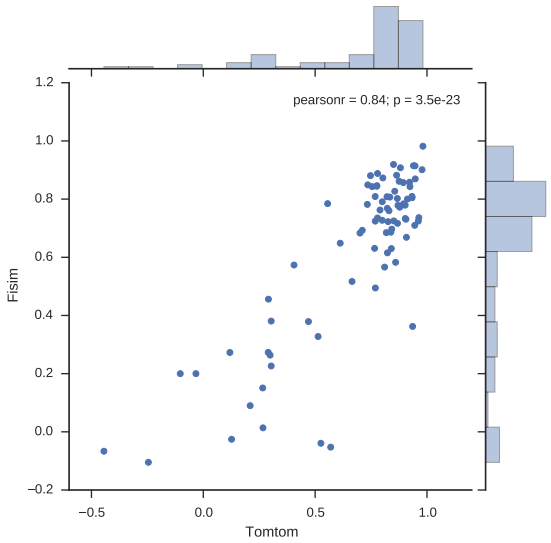


Figure 6. Joint scatter plot and histogram for *gimme roc* and Energy scoring correlation of AUC scores. The two approaches are in agreement and the data is normally distributed

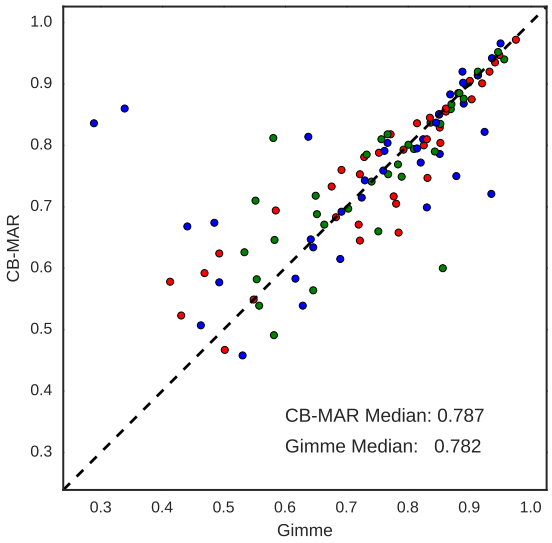


Figure 7. CB-MAR ranking useful in motif discovery: The scatter plot compares the performance of motifs identified by GimmeMotifs and CB-MAR:Tomtom (compare) as evaluated in ChIP-seq data showing the usefulness of data independent approach.

which validates our implementation of energy scoring function.

CB-MAR generates relevant rankings in motif discovery

The first level of application of motif evaluation is in *ab initio* motif discovery, where an algorithm has to narrow down the identified motifs to a few that reflect the binding behaviour of a TF. In addition, the advent of ensemble motif discovery pipelines makes proper motif assessment and ranking even more desirable. By purely using CB-MAR, we were able to correctly identify better or similar motifs (Figure 7) in a majority of the cases. Overall, the quality of motifs identified by GimmeMotifs and CB-MAR are not significantly different (0.97; Wilcoxon rank-sum test) with mean AUC scores of 0.76 and 0.75 respectively. For 6 TFs that had motifs identified by GimmeMotifs being better than CB-MAR motif by more than 0.1 AUC, we checked if choosing the second or third best motif would have any effect on the quality (Figure 8A). We find that choosing the second motif improves the quality in Maz, Yy1, and Atf1, while the third motif is always of a lower quality except for Irf3. For E2f4, the quality is reduced and no effect is observed on Tcfap2 (Ap2). On the other hand, there were 11 TFs with better scores in CB-MAR than GimmeMotifs. When we choose the second or third motif as ranked by GimmeMotifs the quality of the motifs improved for 7 TFs but no effect in E2f1, Ikzf1, Znf263 and Atf3 TFs (Figure 8B).

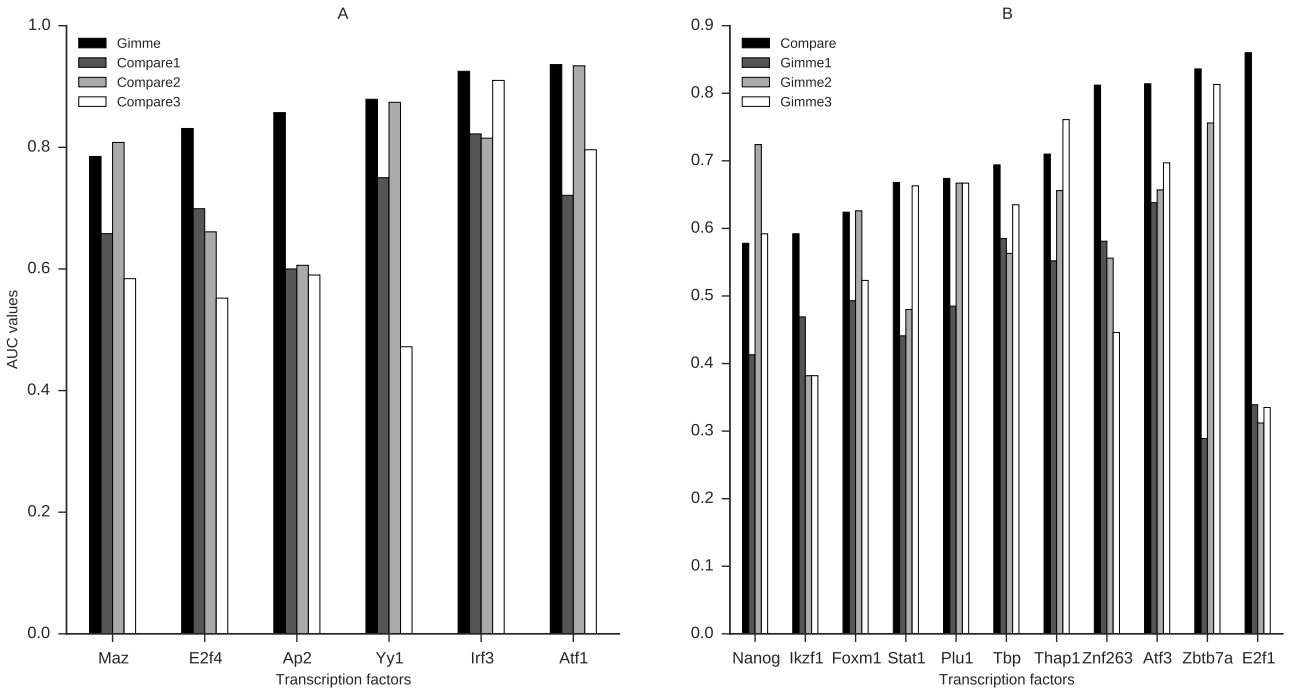


Figure 8. Selecting top three motifs by Gimme and CB-MAR:Tomtom leads to better motifs: The chosen motif is represented by the suffix (Compare1 or Gimme1, for the best and so on). The Y-axis is the mean AUC scores of the motifs in the validation sequences)

Discussion

The number of motifs available for a single TF continues to increase. This offers variability by increasing the binding spectra captured; TFs bind to degenerate sites spread in the genome. However, this is also a challenge. Choosing a binding model is now a daunting task, given that we can already have up to 47 different PWM models in our collection for a single TF generated from a variety of data and algorithms. How can they be ranked to obtain generalized or specific models for a given task? To address this gap, we introduce MARS, a web server that makes PWM motif evaluation and ranking techniques *accessible*, supported by a database of benchmark data and PWM models. How MARS meets Aniba's criteria is highlighted in *italics* – see introduction for details. We ensure MARS can *scale* and *evolve* to support new data and algorithms via the modular design of the algorithms and also by allowing users to upload their own benchmark data. Additionally, CB-MAR allows for evaluation *independence*, and we have demonstrated its *relevance* in motif discovery. Given that there is no agreed standard of motif ranking, we offer the end user a variety of techniques and data for their evaluations to ensure *relevance* to different evaluation conditions.

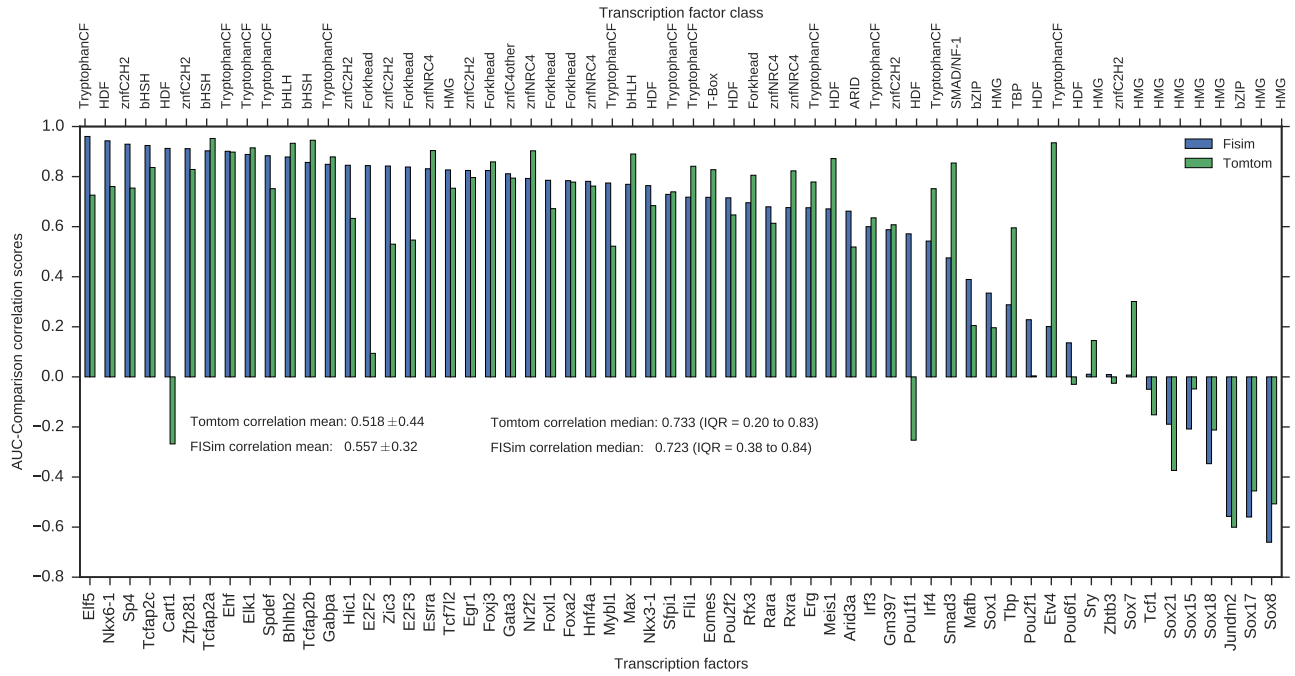


Figure 9. Correlating consistency-based assessment (Tomtom and FISim) with Energy AUC ranking in PBM data: Testing how FISim and Tomtom correlate Energy ranking in PBM data. The TF family class is given on the top axis. The mean ±STD and median with interquartile range statistics are annotated into the figure

Motif quality analysis requires systematic comparative assessment

The TF binding spectra are quite diverse. Therefore, a comparative approach to motif evaluation, using a variety of data and techniques, is necessary in order to understand and make an informed decision on motif quality. For example, we can identify difference in binding behaviour of Zbtb3 TF *in vivo* and *in vitro* by correlating motif performance in PBM and ChIP-seq data (Figure 5); Zbtb3 is known to recognize unmethylated motifs *in vivo* and methylated ones *in vitro* [6]. Furthermore, we can discover that HMG TFs may also bind indirectly, cooperatively or a variation of binding behaviour as captured in PBM data by a low or negative correlation between CB-MAR and energy scoring derived ranks (Figure 9). Indeed HMG TFs, specifically the SOX-related factors, are known to bind cooperatively with partner TFs [17, 19]. In fact, they are believed to form complexes with partner proteins before recognizing the binding site [17]. Therefore, it is expected that the predicted models in PBM data would differ from how they bind *in vivo*. These observations demonstrate the need for a systematic approach to motif quality assessment.

CB-MAR ranks capture *in vivo* binding behaviour

The data-free evaluation approach, CB-MAR, provides a quick and unbiased motif evaluation alternative, especially when the data used in motif discovery is also used for benchmarking. CB-MAR ranks are better correlated with those based on ChIP-seq (R=0.88) than PBM (R=0.73) data, revealing its capability to capture *in vivo* binding (Figures 3A and 9). We further support this argument by using it to successfully identify best PWM models in motif discovery. More details on this in the next section. This implementation reduces the 'reference motifs' bias, an approach in which users considered an algorithm successful if it can predict motifs similar to those in a 'reference database' at a given (usually arbitrary) similarity threshold. The current collections of ChIP-seq and PBM data in our database can only facilitate data-based quality evaluation for less the 300 TFs out of 1352 that have motifs in our database. This demonstrates that this approach is even more desirable.

Between the two motif comparison algorithms tested for CB-MAR, we show that Tomtom captures *in vivo* motif ranks –using ChIP-seq data – better than FISim. FISim is designed to favour similarity of high information or conserved sites [11], revealing that scoring high IC sites better may not match biologically similar motifs. Besides, low information flanking sites have been reported to increases binding specificity in some TFs [10, 16, 21].

CB-MAR ranking useful in motif discovery

The first step after motif discovery is to filter and narrow down to significant motifs. Usually, a partition of the data is held out for testing, but when there is limited data this may not be feasible. Besides, this is only available to the algorithm developers and to motifs generated using sequencing or microarray data (e.g. promoter sequences, ChIP-seq and PBM) and not to those from TF tertiary structures like 3DFootprint [9]. We have demonstrated that the top performing motifs can be identified by CB-MAR in combination with motif clustering to avoid motif duplicates. However, we do not average similar motifs as done by GimmeMotifs. Rather, in a given cluster, the best ranking motif (by similarity to the rest) is chosen. Motif averaging may produce a motif that does not fit biology or reflect TF binding behaviours as demonstrated by the cases where the GimmeMotifs identified motifs performed significantly worse than CB-MAR (Figure 8B) – evaluated against ChIP-seq data.

Conclusions

We have developed MARS, a web server hosting a suite of tools for comparative analysis available from `www.bioinf.ict.ru.ac.za`. This offers choice and flexibility to users since additional test data and motifs can be uploaded, and we do not impose an assessment approach to the users. A major contribution to motif evaluation in this study is the data-independent consistency-based approach (CB-MAR), which offers a good alternative in the absence of benchmark sequence data. We believe that this web server and the algorithms

implemented will help reduce motif redundancy and the continued dependence of low quality 'reference motifs' due to lack of evaluation data. Our suite also acts as a hub for the motifs collated and annotated from various databases and publications.

Acknowledgements

The financial assistance of the South African National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the authors and are not necessarily to be attributed to the NRF. PM funding: NRF/IFR Grant 85362; CK: DST Innovation Doctoral Scholarship Grant 89071. Finally, we acknowledge the ENCODE Consortium for making the data available.

References

1. M. R. Aniba, O. Poch, and J. D. Thompson. Issues in bioinformatics benchmarking: The case study of multiple sequence alignment. *Nucleic Acids Research*, 38(21):7353–7363, 2010.

2. G. Badis, M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, H. Kuznetsov, C.-F. Wang, D. Coburn, D. E. Newburger, Q. Morris, T. R. Hughes, and M. L. Bulyk. Diversity and complexity in DNA recognition by transcription factors. *Science*, 324(5935):1720–1723, June 2009.

3. T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble. MEME Suite: tools for motif discovery and searching. *Nucleic Acids Research*, 37(suppl 2):W202–W208, 2009.

4. T. L. Bailey and P. Machanick. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Research*, 40(17):1–10, Sept. 2012.

5. M. F. Berger, A. a. Philippakis, A. M. Qureshi, F. S. He, P. W. Estep, and M. L. Bulyk. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature biotechnology*, 24(11):1429–35, dec 2006.

6. A. Blattler, L. Yao, Y. Wang, Z. Ye, V. X. Jin, and P. J. Farnham. ZBTB33 binds unmethylated regions of the genome associated with actively expressed genes. *Epigenetics & Chromatin*, 6(1):13, 2013.

7. J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12):1213–8, 2013.

8. X. Chen, T. R. Hughes, and Q. Morris. RankMotif++: a motif-search algorithm that accounts for relative ranks of K-mers in binding transcription factors. *Bioinformatics*, 23(13):i72–i79, July 2007.

9. B. Contreras-Moreira. 3D-footprint: a database for the structural analysis of protein-DNA complexes. *Nucleic Acids Research*, 38(Database):D91–D97, 2010.

10. I. Dror, R. Rohs, and Y. Mandel-Gutfreund. How motif environment influences transcription factor search dynamics: Finding a needle in a haystack. *BioEssays*, 38(7):605–612, 2016.

11. F. Garcia, F. J. Lopez, C. Cano, and A. Blanco. FISim: a new similarity measure between transcription factor binding sites based on the fuzzy integral. *BMC Bioinformatics*, 10:224, 2009.

12. S. Gupta and J. Stamatoyannopoulos. Quantifying similarity between motifs. *Genome Biology*, 8(24), 2007.

13. S. Hannenhalli. Eukaryotic transcription factor binding sites–modeling and integrative search methods. *Bioinformatics*, 24(11):1325–31, June 2008.

14. S. Iantorno, K. Gori, N. Goldman, M. Gil, and C. Dessimoz. Who watches the watchmen? An appraisal of benchmarks for multiple sequence alignment. *Methods in Molecular Biology*, 1079:59–73, 2014.

15. D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science (New York, N.Y.)*, 316(5830):1497–502, 2007.

16. A. Jolma, J. Yan, T. Whitington, J. Toivonen, K. R. Nitta, P. Rastas, E. Morgunova, M. Enge, M. Taipale, G. Wei, K. Palin, J. M. Vaquerizas, R. Vincentelli, N. M. Luscombe, T. R. Hughes, P. Lemaire, E. Ukkonen, T. Kivioja, and J. Taipale. DNA-binding specificities of human transcription factors. *Cell*, 152(1-2):327–339, Jan. 2013.

17. Y. Kamachi and H. Kondoh. Sox proteins: regulators of cell fate specification and differentiation. *Development*, 140(20):4129–44, 2013.

18. C. K. Kibet and P. Machanick. Transcription factor motif quality assessment requires systematic comparative analysis [version 2; referees: 2 approved]. *F1000Research*, 4(ISCB Comm J):1429, Mar. 2016.

19. H. Kondoh and Y. Kamachi. SOX-partner code for cell specification: Regulatory target selection and underlying molecular mechanisms. *International Journal of Biochemistry and Cell Biology*, 42(3):391–399, 2010.

20. T. Lassmann and E. L. L. Sonnhammer. Automatic assessment of alignment quality. *Nucleic Acids Research*, 33(22):7120–7128, 2005.

21. M. Levo, E. Zalckvar, E. Sharon, A. C. Dantas Machado, Y. Kalma, M. Lotam-Pompan, A. Weinberger, Z. Yakhini, R. Rohs, and E. Segal. Unraveling determinants of transcription factor binding outside the core binding site. *Genome Research*, pages 1018–1029, 2015.

22. D. E. Newburger and M. L. Bulyk. UniPROBE: An online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Research*, 37(SUPPL. 1):77–82, Jan. 2009.

23. Y. Orenstein, C. Linhart, and R. Shamir. Assessment of algorithms for inferring positional weight matrix motifs of transcription factor binding sites using protein binding microarray data. *PLoS ONE*, 7(9):e46145, Jan. 2012.

24. Y. Orenstein, E. Mick, and R. Shamir. RAP: accurate and fast motif finding based on protein-binding microarray data. *Journal of computational biology*, 20(5):375–82, May 2013.

25. M. T. Pervez, M. E. Babar, A. Nadeem, M. Aslam, A. R. Awan, N. Aslam, T. Hussain, N. Naveed, S. Qadri, U. Waheed, and M. Shoaib. Evaluating the Accuracy and Efficiency of Multiple-Sequence Alignment Methods. *Evolutionary Bioinformatics Online*, pages 205–217, 2014.

26. D. Quest, K. Dempsey, M. Shafiullah, D. Bastola, and H. Ali. A parallel architecture for regulatory motif algorithm assessment. *2008 IEEE International Symposium on Parallel and Distributed Processing*, pages 1–8, Apr. 2008.

27. J. A. Reuter, D. V. Spacek, and M. P. Snyder. High-Throughput Sequencing Technologies. *Molecular Cell*, 58(4):586–597, 2015.

28. H. S. Rhee and B. F. Pugh. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, 147(6):1408–1419, 2011.

29. M. Safran, I. Dalah, J. Alexander, N. Rosen, T. Iny Stein, M. Shmoish, N. Nativ, I. Bahir, T. Doniger, H. Krug, A. Sirota-Madi, T. Olender, Y. Golan, G. Stelzer, A. Harel, and D. Lancet. Genecards version 3: the human gene integrator. *Database*, 2010, 2010.

30. G. K. Sandve, O. Abul, V. Walseng, and F. Drabløs. Improved benchmarks for computational motif discovery. *BMC Bioinformatics*, 8:193, Jan. 2007.

31. L. Song and G. E. Crawford. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols*, 2010(2):pdb.prot5384, Feb. 2010.

32. The ENCODE Project Consortium, I. Dunham, A. Kundaje, S. F. Aldred, P. J. Collins, C. a. Davis, F. Doyle, C. B. Epstein, S. Frietze, J. Harrow, R. Kaul, J. Khatun, B. R. Lajoie, S. G. Landt, B.-K. B.-K. Lee, F. Pauli, K. R. Rosenbloom, P. Sabo, A. Safi, A. Sanyal, N. Shoresh, J. M. Simon, L. Song, N. D. Trinklein, R. C. Altshuler, E. Birney, J. B. Brown, C. Cheng, S. Djebali, X. Dong, J. Ernst, T. S. Furey, M. Gerstein, B. Giardine, M. Greven, R. C. Hardison, R. S. Harris, J. Herrero, M. M. Hoffman, S. Iyer, M. Kelllis, P. Kheradpour, T. Lassman, Q. Li, X. Lin, G. K. Marinov, A. Merkel, A. Mortazavi, S. C. J. S. L. Parker, T. E. Reddy, J. Rozowsky, F. Schlesinger, R. E. Thurman, J. Wang, L. D. Ward, T. W. Whitfield, S. P. Wilder, W. Wu, H. S. Xi, K. Y. Yip, J. Zhuang, B. E. Bernstein, E. D. Green, C. Gunter, M. Snyder, M. J. Pazin, R. F. Lowdon, L. a. L. Dillon, L. B. Adams, C. J. Kelly, J. Zhang, J. R. Wexler, P. J. Good, E. a. Feingold, G. E. Crawford, J. Dekker, L. Elinitski, P. J. Farnham, M. C. Giddings, T. R. Gingeras, R. Guigó, T. J. T. J. Hubbard, M. Kellis, W. J. Kent, J. D. Lieb, E. H. Margulies, R. M. Myers, J. a. Starnatoyannopoulos, S. a. Tennebaum, Z. Weng, K. P. White, B. Wold, Y. Yu, J. Wrobel, B. a. Risk, H. P. Gunawardena, H. C. Kuiper, C. W. Maier, L. Xie, X. Chen, T. S. Mikkelsen, S. Gillespie, A. Goren, O. Ram, X. Zhang, L. Wang, R. Issner, M. J. Coyne, T. Durham, M. Ku, T. Truong, M. L. Eaton, A. Dobin, T. Lassmann, A. Tanzer, J. Lagarde, W. Lin, C. Xue, B. a. Williams, C. Zaleski, M. Röder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, P. Batut, I. Bell, K. Bell, S. Chakrabortty, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, G. Li, O. J. Luo, E. Park, J. B. Preall, K. Presaud, P. Ribeca, D. Robyr, X. Ruan, M. Sammeth, K. S. Sandu, L. Schaeffer, L.-H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. H. Wang, Y. Hayashizaki, A. Reymond, S. E. Antonarakis, G. J. Hannon, Y. Ruan, P. Carninci, C. a. Sloan, K. Learned, V. S. Malladi, M. C. Wong, G. P. Barber, M. S. Cline, T. R. Dreszer, S. G. Heitner, D. Karolchik, V. M. Kirkup, L. R. Meyer, J. C. Long, M. Maddren, B. J. Raney, L. L. Grasfeder, P. G. Giresi, A. Battenhouse, N. C. Sheffield, K. a. Showers, D. London, A. a. Bhinge, C. Shestak, M. R. Schaner, S. K. Kim, Z. Z. Z. Zhang, P. a. Mieczkowski, J. O. Mieczkowska, Z. Liu, R. M. McDaniell, Y. Ni, N. U. Rashid, M. J. Kim, S. Adar, T. Wang, D. Winter, D. Keefe, V. R. Iyer, K. S. Sandhu, M. Zheng, P. Wang, J. Gertz, J. Vielmetter, E. C. Partridge, K. E. Varley, C. Gasper, A. Bansal, S. Pepke, P. Jain, H. Amrhein, K. M. Bowling, M. Anaya, M. K. Cross, M. a. Muratet, K. M. Newberry, K. McCue, A. S. Nesmith, K. I. Fisher-Aylor, B. Pusey, G. DeSalvo, S. S. Balasubramanian, N. S. Davis, S. K. Meadows, T. Eggleston, J. S. Newberry, S. E. Levy, D. M. Absher, W. H. Wong, M. J. Blow, A. Visel, L. a. Pennachio, L. Elnitski, H. M. Petrykowska, A. Abyzov, B. Aken, D. Barrell, G. Barson, A. Berry,

A. Bignell, V. Boychenko, G. Bussotti, C. Davidson, G. Despacio-Reyes, M. Diekhans, I. Ezkurdia, A. Frankish, J. Gilbert, J. M. Gonzalez, E. Griffiths, R. Harte, D. a. Hendrix, T. Hunt, I. Jungreis, M. Kay, E. Khurana, J. Leng, M. F. Lin, J. Loveland, Z. Lu, D. Manthravadi, M. Mariotti, J. Mudge, G. Mukherjee, C. Notredame, B. Pei, J. M. Rodriguez, G. Saunders, A. Sboner, S. Searle, C. Sisu, C. Snow, C. Steward, E. Tapanari, M. L. Tress, M. J. van Baren, S. Washieti, L. Wilming, A. Zadissa, Z. Zhengdong, M. Brent, D. Haussler, A. Valencia, A. Raymond, N. Addleman, R. P. Alexander, R. K. Auerbach, K. Bettinger, N. Bhardwaj, A. P. Boyle, A. R. Cao, P. Cayting, A. Charos, Y. Cheng, C. Eastman, G. Euskirchen, J. D. Fleming, F. Grubert, L. Habegger, M. Hariharan, A. Harmanci, S. Iyenger, V. X. Jin, K. J. Karczewski, M. Kasowski, P. Lacroute, H. Lam, N. Larnarre-Vincent, J. Lian, M. Lindahl-Allen, R. Min, B. Miotto, H. Monahan, Z. Moqtaderi, X. J. Mu, H. O'Geen, Z. Ouyang, D. Patacsil, D. Raha, L. Ramirez, B. Reed, M. Shi, T. Slifer, H. Witt, L. Wu, X. Xu, K.-K. Yan, X. Yang, K. Struhl, S. M. Weissman, S. a. Tenebaum, L. O. Penalva, S. Karmakar, R. R. Bhanvadia, A. Choudhury, M. Domanus, L. Ma, J. Moran, A. Victorsen, T. Auer, L. Centarin, M. Eichenlaub, F. Gruhl, S. Heerman, B. Hoeckendorf, D. Inoue, T. Kellner, S. Kirchmaier, C. Mueller, R. Reinhardt, L. Schertel, S. Schneider, R. Sinn, B. Wittbrodt, J. Wittbrodt, G. Jain, G. Balasundaram, D. L. Bates, R. Byron, T. K. Canfield, M. J. Diegel, D. Dunn, A. K. Ebersol, T. Frum, K. Garg, E. Gist, R. S. Hansen, L. Boatman, E. Haugen, R. Humbert, A. K. Johnson, E. M. Johnson, T. M. Kutyavin, K. Lee, D. Lotakis, M. T. Maurano, S. J. Neph, F. V. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds,. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

33. M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. a. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Régnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, 23(1):137–44, Jan. 2005.

34. S. J. van Heeringen and G. J. C. Veenstra. GimmeMotifs: a de novo motif prediction pipeline for ChIP-sequencing experiments. *Bioinformatics*, 27(2):270–271, 2011.

35. M. Vihinen. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC Genomics*, 13(Suppl 4):S2, 2012.

36. M. T. Weirauch, A. Cote, R. Norel, M. Annala, Y. Zhao, T. R. Riley, J. Saez-Rodriguez, T. Cokelaer, A. Vedenko, S. Talukder, H. J. Bussemaker, Q. D. Morris, M. L. Bulyk, G. Stolovitzky, and T. R.

Hughes. Evaluation of methods for modeling transcription factor sequence specificity. *Nature Biotechnology*, 31(2):126–34, Feb. 2013.

37. E. G. Wilbanks and M. T. Facciotti. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS ONE*, 5(7):e11471, Jan. 2010.

38. E. Wingender, T. Schoeps, and J. Dönitz. TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Research*, 41(Database issue):D165–70, Jan. 2013.

39. F. Zambelli, G. Pesole, and G. Pavesi. Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Briefings in bioinformatics*, 14(2):225–37, Mar. 2013.

40. Y. Zhao, D. Granas, and G. D. Stormo. Inferring binding energies from selected binding sites. *PLoS Computational Biology*, 5(12):e1000590, Dec. 2009.

41. S. Zhong, X. He, and Z. Bar-Joseph. Predicting tissue specific transcription factor binding sites. *BMC Genomics*, 14:796, Jan. 2013.