RESEARCH ARTICLE

REVISED Transcription factor motif quality assessment requires systematic comparative analysis [version 2; referees: 2 approved]

Caleb Kipkurui Kibet, Philip Machanick

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

v2 **First published:** 11 Dec 2015, **4**(ISCB Comm J):1429 (doi: 10.12688/f1000research.7408.1)

Latest published: 14 Mar 2016, **4**(ISCB Comm J):1429 (doi: 10.12688/f1000research.7408.2)

Abstract

Transcription factor (TF) binding site prediction remains a challenge in gene regulatory research due to degeneracy and potential variability in binding sites in the genome. Dozens of algorithms designed to learn binding models (motifs) have generated many motifs available in research papers with a subset making it to databases like JASPAR, UniPROBE and Transfac. The presence of many versions of motifs from the various databases for a single TF and the lack of a standardized assessment technique makes it difficult for biologists to make an appropriate choice of binding model and for algorithm developers to benchmark, test and improve on their models. In this study, we review and evaluate the approaches in use, highlight differences and demonstrate the difficulty of defining a standardized motif assessment approach. We review scoring functions, motif length, test data and the type of performance metrics used in prior studies as some of the factors that influence the outcome of a motif assessment. We show that the scoring functions and statistics used in motif assessment influence ranking of motifs in a TF-specific manner. We also show that TF binding specificity can vary by source of genomic binding data. We also demonstrate that information content of a motif is not in isolation a measure of motif quality but is influenced by TF binding behaviour. We conclude that there is a need for an easy-to-use tool that presents all available evidence for a comparative analysis.

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	2
REVISED version 2 published 14 Mar 2016	✓ report	✓ report
version 1 published 11 Dec 2015	? report	? report

1 **Trevor W. Siggers**, Boston University USA

2 **Jan Grau**, Martin Luther University of Halle-Wittenberg Germany

Discuss this article

Comments (0)

This article is included in the ISCB Africa ASBCB Conference on Bioinformatics channel.

Corresponding authors: Caleb Kipkurui Kibet (calebkibet88@gmail.com), Philip Machanick (p.machanick@ru.ac.za)

How to cite this article: Kibet CK and Machanick P. **Transcription factor motif quality assessment requires systematic comparative analysis [version 2; referees: 2 approved]** *F1000Research* 2016, **4**(ISCB Comm J):1429 (doi: 10.12688/f1000research.7408.2)

Grant information: 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.

Competing interests: No competing interests were disclosed.

First published: 11 Dec 2015, **4**(ISCB Comm J):1429 (doi: 10.12688/f1000research.7408.1)

REVISED **Amendments from Version 1**

We thank the reviewers for their comments. The paper has been updated in response to their comments as follows:

1. A sentence is added in the background section to update information on the performance of PWM models.
2. Most of the figures have been updated or modified, especially the captions. Figure 6 was added while Figure 5 was replaced.
3. We include analysis on the effect of negative sequences on motif ranking. We added a subsection, 'How choice of negative (background) sequence affects motif ranking' and Figure 6.
4. We include data on a re-run of the analysis using PBM data. A paragraph has been included in the Data subsection (2nd paragraph), a new subsection, 'Effect of PBM data on motif assessment' in the results section discussed in the discussion section.
5. A subsection of the results section replaced with 'Effect of statistics on motif ranking'. This goes with the new Figure 5. In addition, Table 2, which offered supporting information to the old Figure 5, has also been removed.
6. The notations used in the formulas have been harmonized.
7. We have added a definition of MNCP to the 'Statistical measures of performance' subsection of the methodology section.
8. A supplementary section containing Supplementary Figure 1 to Supplementary Figure 4 for PBM data, for comparison with the equivalent ChIP-seq Figure 7 to Figure 10, has been added.
9. We have added a list of ENCODE ChIP-seq data used to the repository as Supplementary Table 3 (Table S3).
10. Finally, we have made other minor changes in response to the reviewers comments, as noted in our detailed response to the reviews.

See referee reports

Background

Understanding gene regulation remains a long-standing problem in biological research. The main players, transcription factors (TFs), are proteins that bind to short and potentially degenerate sequence patterns (motifs) at gene regulatory sites to promote or repress expression of target genes. The search for a code to predict binding sites and model binding affinity of TFs has led to several experimental techniques and motif discovery algorithms being developed (Figure 1).

A position weight matrix (PWM) is the common form of representing TF binding specificity. For a motif of length L, the corresponding PWM is a $4 \times L$ matrix of probabilities of observing a base b (A, C, G or T) at position i through L. Other forms of representing TF binding specificity have been introduced[1–4], but Weirauch *et al.* showed that a well-trained PWM performs comparably to some of the above well trained complex models[5]. However, recent studies[6–8] have reported significant improvement to the PWM by models that consider nucleotide inter-dependencies. The persistent popularity of PWM can be attributed to its simplicity and ease of use as well as the ease of visualizing a PWM using a sequence logo[9]. Motifs can be found using a variety of methods including algorithms that do *de novo* motif discovery from sequences containing binding sites[10–12] and *in vitro* methods such as protein binding microarrays (PBM)[13] and high-throughput systematic evolution of ligands by exponential enrichment (HT-SELEX)[14].

Initially, the low resolution of the available experimental techniques for TF binding specificity detection was a hindrance to the quality of binding models. However, next generation sequencing and techniques like chromatin immunoprecipitation (ChIP) followed by deep sequencing (ChIP-seq)[15] and exonuclease cleavage in ChIP-exo[16] that measure TF *in vivo* occupancy, have improved the resolution to single-nucleotide level. In addition to providing high resolution data for motif discovery, they are a useful resource to test the quality of the available motifs since they are TF specific. However, no benchmark capable of assessing the growing range of published motifs is available and quality measures are largely subjective[17].

Although it is possible that PWM models' ability to describe TF binding may be getting saturated, the lack of a robust approach to test the quality of a model and maximize the best-performing ones may also be hampering improvement in performance. How are the algorithms being developed, tested and improved? Furthermore, the number of motif finding algorithms from dissimilar data sets and subsequently the number of motif models for a single TF generated, continue to increase. There are at least 44 PWM motif models available in 14 different databases for Hnf4a alone. How does the end-user decide which motif to use? In this study, we review and test the approaches used to evaluate TF binding models.

Review of motif assessment approaches

The available motif assessment techniques can be divided into three categories: assess by binding site prediction, motif comparison or, by sequence scoring and classification.

Binding site prediction

Early review and assessment of motif-finding algorithms tested tools on the ability to predict sites of motifs, known or inserted into the sequence. Tompa *et al.* tested motif discovery algorithms by their ability to predict sites of inserted motifs using statistical measures for site sensitivity and correlation coefficient[18]. In this first comprehensive study, they found that a motif assessment problem is complex and admitted that inserting random motifs fails to capture the biological condition of TF binding. Later, Hu *et al.*[19] used real RegulonDB binding data in a large-scale analysis of five motif-finding algorithms. The tools available at that time performed poorly – "15–25% accuracy at the nucleotide level and 25–35% at the binding site level for sequences of 400 nt long" – largely due to the poor quality of RegulonDB annotations[20].

Sandve and colleagues[21–23] tested motif discovery algorithms using sequences with real and inserted binding sites as benchmarks; from Transfac, and the third-order Markov model respectively. Quest and colleagues[24] developed the Motif Tool Assessment Platform (MTAP) as an automated test of motif discovery tools. However, this was computationally expensive and was made obsolete by new experimental data and algorithms.

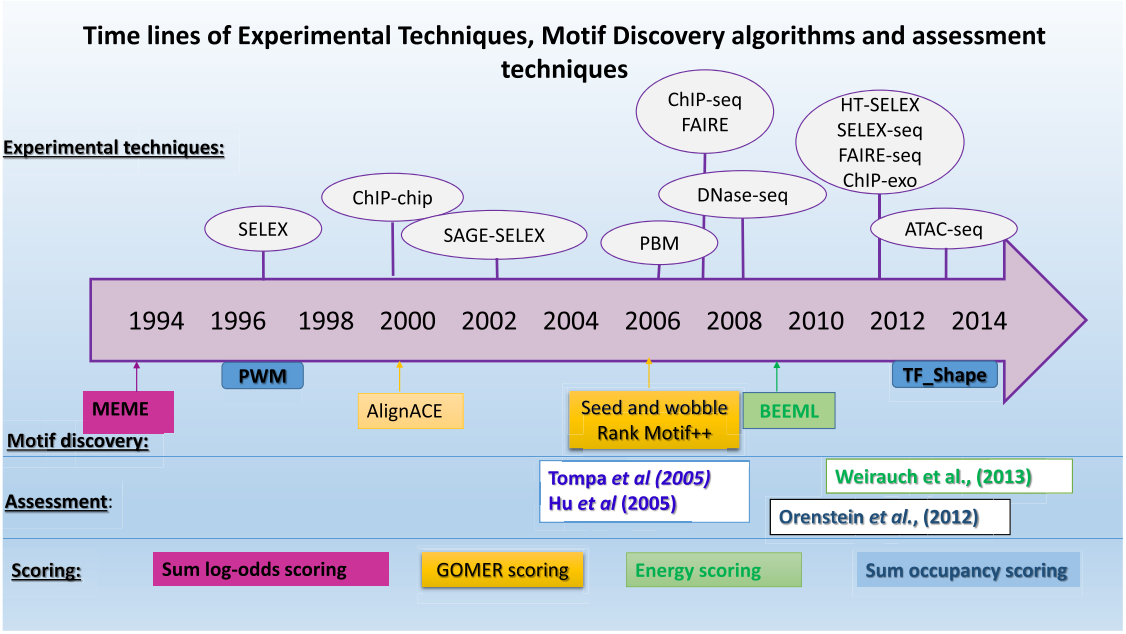

Figure 1. Evolution of motif scoring functions with experimental techniques and algorithms. Tompa *et al.*[18] and Hu *et al.*[19] assessed the motifs by binding site prediction while Orenstein *et al.*[28] and Weirauch *et al.*[5] used scoring. The scoring techniques are colour coded for the motif discovery or assessment where they were used.

The most comprehensive assessment based on binding site prediction so far has been by the Regulatory Sequence Analysis Tools (RSAT) consortium. In their 'matrix quality' script, they use theoretical – information content (IC) and *E*-values – and empirical scores computed by predicting binding sites in RegulonDB, ChIP-chip and ChIP-seq positive and negative control sequences[20].

Inadequate knowledge of TF binding sites has mainly hampered the ability to assess motifs and algorithms by binding site prediction. Predicting binding sites that are inserted or known in the sequences cannot accurately identify unknown true sites. Techniques that identify such sites may be penalized. Until TF binding sites are well annotated, this technique cannot be confidently utilized.

Motif comparison

Novel motifs can be assessed by comparison to 'reference motifs' using the sum of square deviation, Euclidean distance and other statistics that measure divergence between two PWMs[25,26]. Thomas-Chollier *et al.* proposed a motif comparison approach for their RSAT algorithm where they combine multiple metrics, including Pearson's correlation, width normalized correlation, logo dot product, correlation of IC, normalized Sandelin-Wasserman, sum of squared distances and normalized Euclidean similarity for each matrix pair[27]. They then unified all of these scores to ranks whereby the mean of the ranks is considered the overall score.

Assessing motifs by comparison, as currently implemented, only tests similarity to the available motifs with little information on quality and ranks of the motifs. It assumes accuracy of 'reference motifs', with no way of assessing novel ones. In addition, the definition of 'reference motifs' remains largely subjective.

Assessment by scoring

Motif assessment has since shifted towards scoring positive sequences known to contain binding sites and negative background sequences without binding sites, driven by high-throughput sequencing techniques[5,28–30]. This avoids the need to identify binding sites *a priori* by focusing on the ability to classify the two sets of sequences. The differences in the assessments arise from the choice of sequences to use as positive and negative, the thresholds used to identify binding sites, the length of the sequences in both sets, the scoring function and the statistic used to quantify the performance of the tool.

For ChIP-seq data, the main difference is that the length of sequences (250bp[28]; 600bp[30], 100bp[5] or 60bp[31]) and the choice of negative sets (300bp downstream;[28,30]; random sequences, 5000bp from a transcription start site (TSS) or random genomic sequences[5], or flanking sequences[31]) differ greatly in sequence scoring. In addition Agius *et al.*[31], test PWMs and support vector regression (SVR) models in the 36bp sliding window of the test sequences, a deviation from the rest of the techniques. All these differences, in addition to the scoring functions and statistics used, can lead to variations in the results of comparative analyses. Users and algorithm developers therefore have to frequently re-invent the wheel to test their tools.

Figure 1 shows the evolution of experimental motif discovery assessment techniques. We have not focused on the experimental techniques or motif discovery algorithms as excellent reviews are already available[17,32]. Rather, we focus on TF binding models represented as a PWM and aim to determine how the choice and length of benchmark sequences, scoring functions, and the statistics influence motif assessment. We hope that this study will highlight

some of the pitfalls in the previous motif assessments and provide a starting point for a standard in motif assessment that will ensure comparability and reuse of results.

Methods

Data

Human uniform ChIP-seq data were downloaded from the ENCODE consortium[33] (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform) (List of ENCODE data used available in Table_S3[66]). For each peak file, we used BEDTools v2.17.0[34] to extract the 500 highest scored sequences (only peaks with no repeat masked sequences were used) of 50, 100, and 250bp centred on the ChIP-seq peaks as a positive set. Our choice for top 500 sequences was informed by our understanding of previous research[35,36]; using a TF-specific percentile of the available peaks did not make any significant difference (data not shown). A negative set of a similar number of sequences and length was extracted 500bp downstream from the highest coordinate (highest coordinate + 500) of the positive sequences.

When using PBM data to rank motifs, we mainly adopted the definition of positive and negative sets described by Chen *et al.*[37]. A given motif is used to score a 36bp sequence for each spot using the different scoring functions. For this analysis, we only found nine TFs that had comparable data in ChIP-seq and PBM. These were: Egr1, Esrra, Gata3, Hnf4a, Mafk, Max, Myb, Pou2f2 and Tcf3. The data from Badis *et al.*[38] were downloaded from UNIPROBE database[13]. A detailed Ipython notebook on this analysis can be found in https://github.com/kipkurui/Kibet-F1000Research.

We used motifs from a number of databases and publications listed in Table 1. The TFs used in this analysis were selected based on availability of ChIP-seq data with motifs in at least 10 motif databases. We converted these motifs from their various formats into MEME format and scored the positive and negative sequences with GOMER, occupancy, energy and log-odds scoring functions.

We quantify how each motif performs using AUC, MNCP, Spearman's and Pearson correlation (Figure 2). This was implemented in a Python module which is available free from https://github.com/kipkurui/Kibet-F1000Research. This repository also contains raw data and Ipython notebooks that document how to reproduce the analysis we describe in this paper.

Table 1. Source of motifs used in the analysis. "Source" refers to the experimental technique used to generate the motifs while "mixed" motifs are generated using a variety of techniques. The specific motifs in MEME format used for this analysis are provided in the data repository[66].

Database	Source	Size	Reference
JASPAR	Mixed	127	39
UniPROBE	PBM	386	13
Jolma	HT-SELEX	843	14
Zhao	PBM-BEEML	419	40
POUR	ChIP-seq	292	41
HOCOMOCO	Mixed	426	42
SwissRegulon	Mixed	297	43
TF2DNA	3D Structures	1314	44
HOMER	ChIP-seq	264	45
Chen2008	ChIP-seq	12	35
3DFOOTPRINT	3D Structures	297	46
GUERTIN	ChIP-seq	609	47
CSP-BP	Mixed	734	48
ZLAB	ChIP-seq	409	36

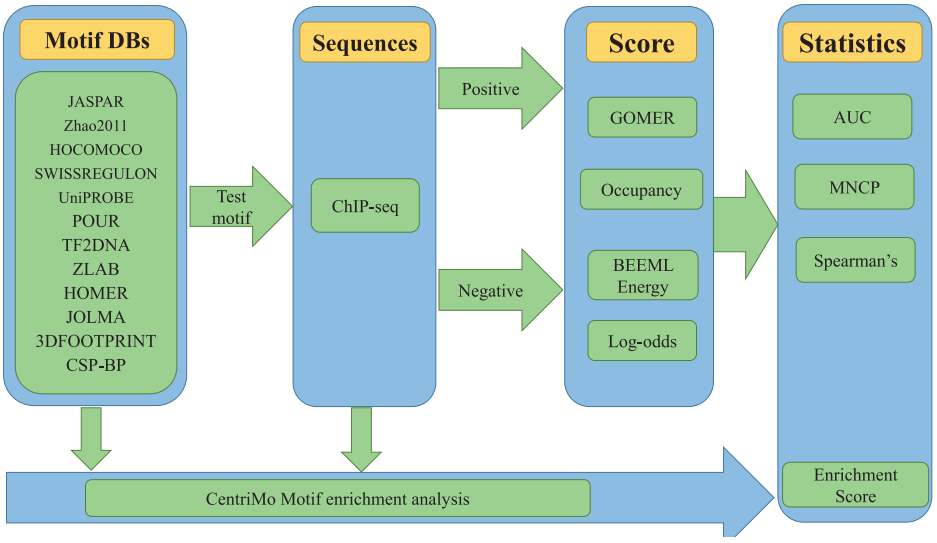

Figure 2. Methodology flow diagram. For a given transcription factor, all motifs available in various databases are extracted and used to score the given test sequences. The motifs are then ranked based on a given statistic.

Scoring functions

When testing motifs by scoring ChIP-seq or PBM data, multiple scoring functions are available, which may affect the outcome. In the section that follows, we describe the scoring functions tested, as well as provide a review of how they have been previously applied.

GOMER scoring

The GOMER scoring framework was introduced by Granek *et al.*[49] but adapted for PBM sequence scoring[37,38]. It seeks to compute the probability $g(S,\Theta) = exp(f(S,\Theta))$ that a TF, given PWM Θ, will bind to at least one of the sub-sequences of S. This assumes that each site can be bound independently

$$g(S,\Theta) = 1 - \prod_{t=1}^{L-k} 1 - P(S_{t:t+k}|\Theta) \quad (1)$$

where L is the length of sequence S, and $S_{t:t+k}$ is the sub-sequence of S from position t to $t+k$ inclusive. See Chen *et al.*[37] for more details.

Occupancy score

The occupancy score calculates the occupancy of a PWM Θ for sub-sequence S^i of length k as the product of the probabilities of each base in S^i using equation 2.

$$f(S^i,\Theta) = \prod_{j=1}^{k} \Theta_j[S^i_j]. \quad (2)$$

For a sequence S of length L, the sum of the occupancies of all sub-sequences S^i (sum occupancy)[28,50], the maximum score (maximum occupancy)[30], or the average occupancy (average motif affinity – AMA) have been used. Sum occupancy is defined in equation 3:

$$f(S,\Theta) = \sum_{t=0}^{L-k} \prod_{j=1}^{k} \Theta_j[S^i_{t+j}]. \quad (3)$$

BEEML-PBM energy scoring

The energy scoring framework of binding energy estimation by maximum likelihood for protein binding microarrays (BEEML-PBM)[4] computes the logarithm of base frequencies with the idea that this is proportional to the energy contributions of the bases. The binding energy at each location is computed; the lower the binding energy, the higher the binding affinity. For each sequence, the sub-sequence with the lowest binding energy represents the score of the sequence. It has mainly been used to score PBM data[5,30].

The probability that sub-sequence S^i is bound is given by equation 4,

$$P(S^i \, is \, bound) = \frac{1}{1 + e^{E(S^i)-\mu}}, \quad (4)$$

where, for a sub-sequence S^i, $E(S^i)$ is given equation 5,

$$E(S^i) = \sum_{b=A}^{T} \sum_{t=1}^{L} \epsilon(b,t) S^i(b,t), \quad (5)$$

for binding site of length L, $\epsilon(b, t)$ is the energy contribution of base b while $S^i(b, t)$ is an indicator function of site t within S^i(1 with base b, 0 otherwise).

Log-odds scoring

In log-odds scoring, used by a majority of the MEME Suite tools[51], the score for a given site is the sum of the log-odds ratios of a PWM at the match site. For a sub-sequence S^i of length L scored using PWM Θ, the log-odds score is given by equation 6,

$$LogOdds(S^i,\Theta,p) = \sum_{t=1}^{L} \sum_{b=A}^{T} S^i(b,t) log \frac{\Theta_{t,b}}{P_b}. \quad (6)$$

where p_b the background probability (uniform background probability of 0.25 is used) and $S^i(b, t)$ is an indicator function of site t as in equation 5.

The score for a given sequence can then be derived by summing individual scores or by finding the maximum score. Sum log-odds scoring has generally been used by MEME Suite tools while maximum log-odds scoring has also been used to compare motifs represented differently (PWM, *k*-mer and SVM models) against one another[30,31]. Each of these approaches has inherent advantages but may produce inconsistent results.

Statistical measures of performance

With the scores of each motif for the sequences acquired, binding prediction can be evaluated by various statistics. The area under the receiver operating characteristic curve (AUC)[52] has been widely used, especially with the advent of PBM[5,28,37]. In addition to popularizing AUC, Clarke *et al.*[52] also introduced a novel metric, mean normalized conditional probability (MNCP), for quantifying the correlation between DNA features and gene regulation. This statistic has been applied for motif assessment in GIMME motifs[53] and is said to be less affected by the presence of false positives compared with AUC since it places emphasis on true positives. MNCP is a rank-based statistic that determines if mean occurrence of a motif in test sequences is higher than the mean occurrence in a random set. Each set of sequences is ranked based on the mean occurrence, and the MNCP calculated by finding the mean of the normalized ratio of the two sets of ranks. We use MNCP to test how it contributes to better prediction in an effort to encourage its use.

Pearson and Spearman's rank correlation are still widely used as a measure of motif performance. Spearman's rank correlation has been used for PBM and ChIP-seq sequences[28] while Pearson's correlation was used by Weirauch *et al.*[5]. However, Weirauch *et al.* cautioned on the use of Spearman's correlation for PBM data citing its inability to exclude low intensity probes. We wish to check the usefulness of correlation statistics in motif assessment.

In addition to comparing the scoring approaches, we use CentriMo version 4.10.0 in differential mode[54] – an option that tests differences in motif enrichment between two sequence sets – in a novel way for motif assessment. We set differential mode parameters for local rather than central enrichment of all the input motifs in the positive (primary) and negative (control) set, as described in the Data section, by using a very large threshold. The negative log of

the *E*-value is used as the measure of a motif's enrichment and rank. Motif enrichment analysis has previously been performed[36] using the FIMO algorithm[55] to scan for motif matches in sequences and calculate an enrichment value.

Results

Length of sequences has a little effect on motif performance

The size of the putative binding region – length of the sequences in each data set – is to some extent a proxy for how accurate the ChIP-seq experiment was. If the result was accurate a narrow region should contain the true site. For the three variants of sequence length, we did not observe a significant effect (p=0.113, for 50 and 100; p=0.0545, 50 and 250; p=0.678, 100 and 250bp – Wilcoxon rank-sum test) on the scoring of the sequences (Figure 3). The scores assigned for each sequence length, however, seems to indicate how the TFs bind. Motifs with higher scores at lower sequence length (50 or 250bp) are generally enriched at the ChIP-seq peak, which is also a strong indicator of direct binding[56]. This is consistent with a previous observation that a successful ChIP-seq experiment localizes binding within about 100bp of the true site[57]. Others with significantly better AUC values at 250bp sequence length like Elf1 (p=0.017, Wilcoxon rank-sum test) and Sp1 (p=0.013, Wilcoxon rank-sum test)[58], are known to bind cooperatively.

Tissue or cell line of the data could affect enrichment

Transcription factors bind to their possible sites in a sequence-specific manner. Some actually have alternative binding motifs depending on the tissue or cell line. Unless the purpose of motif assessment is to identify tissue-specific binding, if data is available from more than one cell line, an average of the scores should be used. For example, in Figure 4, we show that the rank correlation of the motif scores in different cell lines can be as low as 0.8 for GOMER scoring (or as low as 0.65 when energy scoring is used), and not 1 or very close to 1 as would be expected if the cell line had no effect. In addition, FOXA1_1.GUERTIN motif is differentially enriched only in the A549 cell line (although this could be an outlier).

In light of this possible effect, the results displayed throughout this paper are based on the mean score of all the available ChIP-seq data sets to avoid a bias towards cell line-specific motifs.

How choice of negative (background) sequences affect motif ranking

In motif discovery, the choice of background sequences has significant effects on the motifs identified. We sought, therefore, to test whether motif scoring would be affected in a similar way. In addition to downstream sequences, we used a dinucleotide shuffled set from the positive sequences. The scores obtained using dinucleotide shuffled positive sequences were always lower than those for downstream sequences. We then computed and plotted the rank correlation of scores normalized by maximum score for each TF, from which we find that it affects the ranks of the motifs (Figure 5) in a TF-specific manner. However, the scores from the two sets of negative sequences used are not significantly different (p=0.484, Wilcoxon rank-sum test). For Myb, the low correlation could be attributed to how it binds, indirectly[59].

Effect of statistic on motif ranking

The statistic used, whether it measures scores correlation or ability to classify the two sets of sequences, will definitely have an effect on how we interpret the results of the analysis. Generally, the motifs ranks based on AUC and MNCP statistics' are not significantly different (p=0.52, Wilcoxon rank-sum test), but the ranks based on

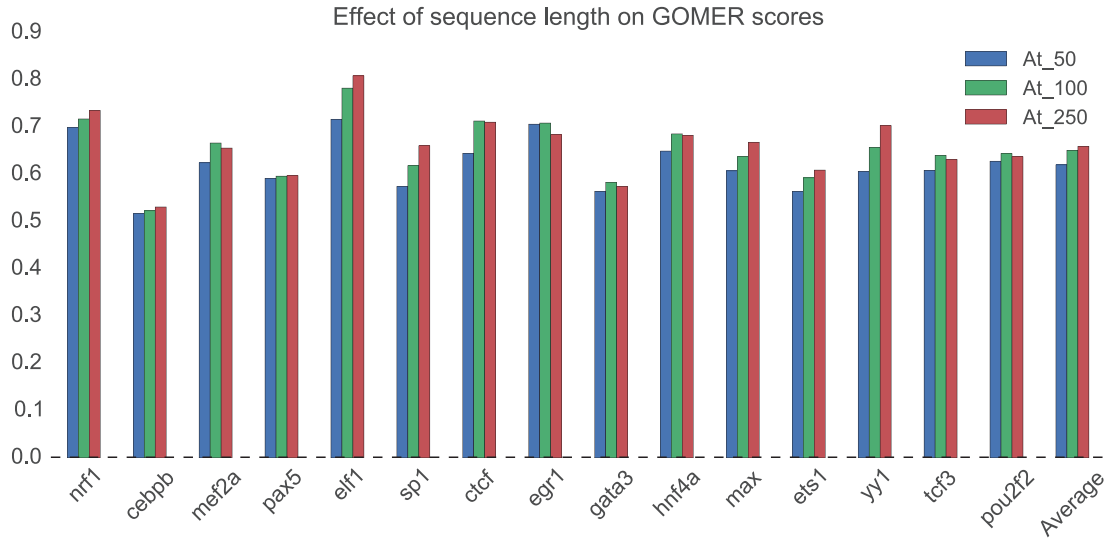

Figure 3. Effect of sequence length. Using all the motifs for each of the 15 TFs, we tested the effect of sequence length (50bp, 100bp and 250bp) using GOMER scoring on ChIP-seq data. For each TF, the mean of the AUC of the motifs is computed and the mean of all the 15 TFs computed to obtain the average. The motifs used in this analysis are available as supplementary data in the repository[66].

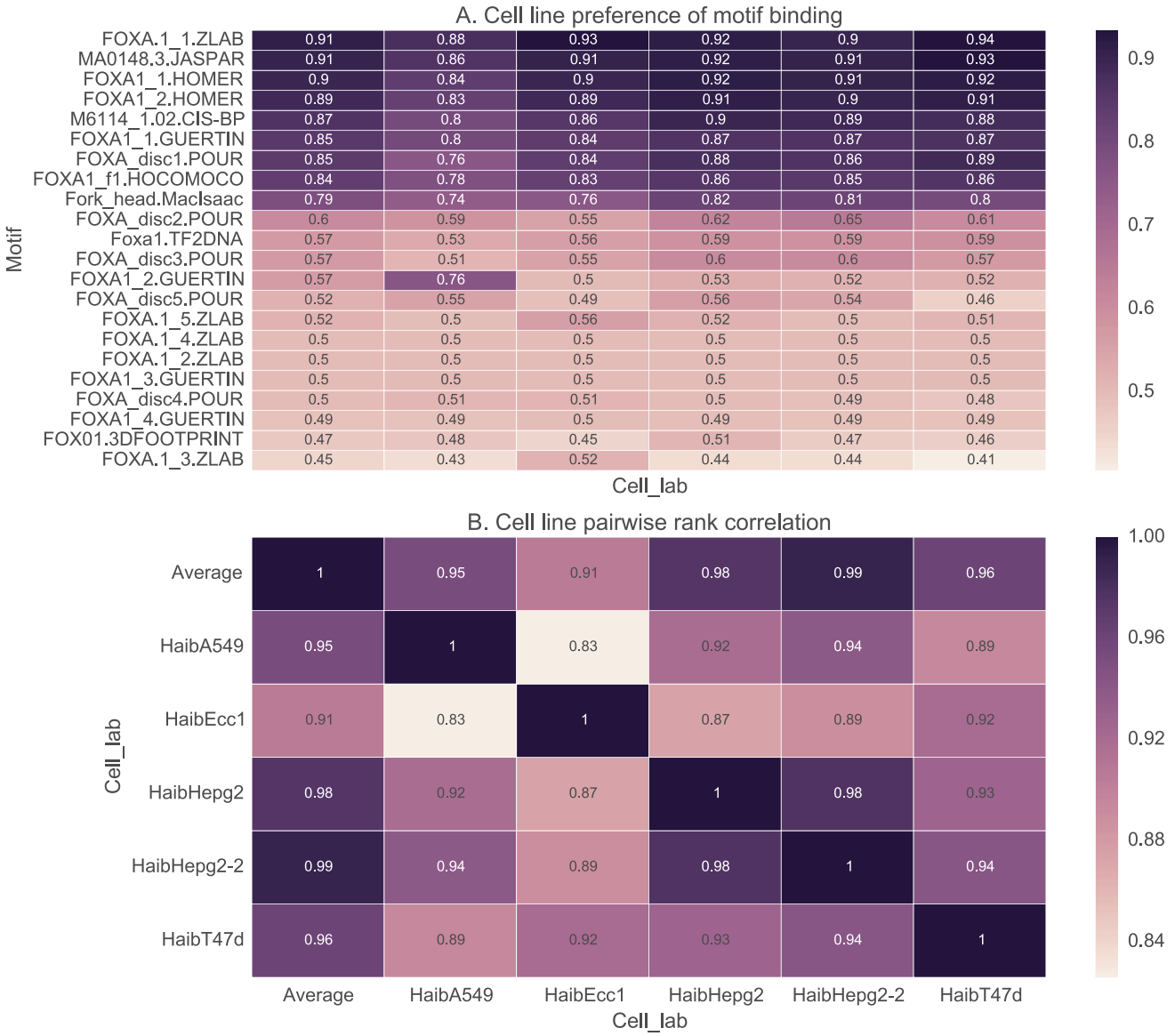

Figure 4. Cell line-specific binding. In **A**, we show how the ranks of the motifs can be influenced by the cell line used in the analysis. Foxa motifs are used to score each of the five cell lines using GOMER scoring and quantified with AUC values. Similar results are obtained with other scoring functions. In **B**, we show how the ranks assigned to the motifs are correlated among the cell lines.

Pearson and Spearman's differ significantly from MNCP or AUC scores (p=0.006 and 0.002 respectively, Wilcoxon rank-sum test). The large standard deviations of the correlation statistics' scores, as shown by the error bars in the Figure 6, shows how unreliable the use of correlation statistics to rank the motifs can be. The correlation scores are also quite low.

Effect of scoring function is transcription factor specific

We tested the ability of PWM models to discriminate positive (top 500 peaks of width 100bp centred on the peak) and negative (500 peaks 100bp wide located 500bp downstream from the positive) sequence sets using five scoring functions. Maximum and sum log-odds scoring had low discriminative power for most of the motifs when AUC (Figure 7) and MNCP (Figure 8) statistical measures are used. However, sum log-odds scoring had some good performance (over 0.55 AUC scores) for some TF motifs like Max, Nrf1, Tcf3 and Pax5.

There is no significant difference in performance when GOMER, energy or occupancy scores (sum, maximum and AMA) are used for scoring (Figure 7B) with AUC statistic (see Table_S1 for details of statistical significance). Also, we did not observe any significant

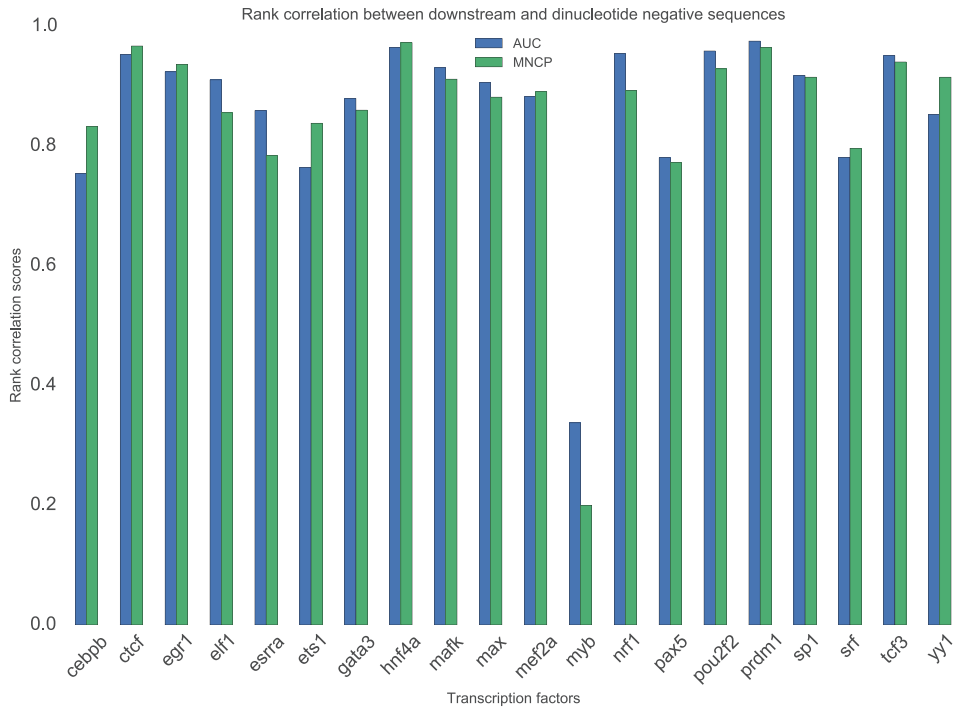

Figure 5. Influence of negative sequences on motif ranking. For each TF, the available motifs are used to score positive and two sets of negative sequence; downstream set and a dinucleotide shuffle of the positive set (see text for details). The figure displays a rank correlation of normalized MNCP and AUC scores from the two sets of negative sequences. Pearson and Spearman's correlation do not require negative sequences therefore, they are not affected. We only show results based on GOMER scoring, but similar conclusions can be made from the other scoring functions.

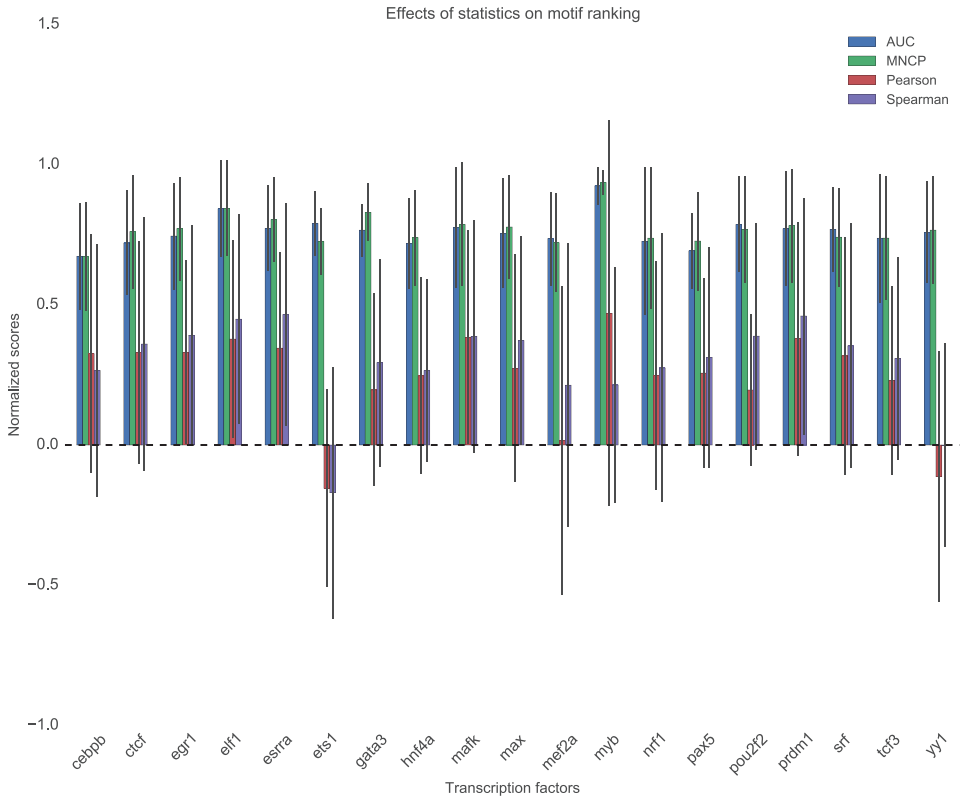

Figure 6. Effects of statistics on motif ranking. For each TF, the motifs are used to score sequences using GOMER scoring function and ranks determined by MNCP, AUC, Pearson and Spearman's rank correlation. In this figure, we compute the mean normalized scores and compute the standard deviation for each TF, which is displayed as error bars.

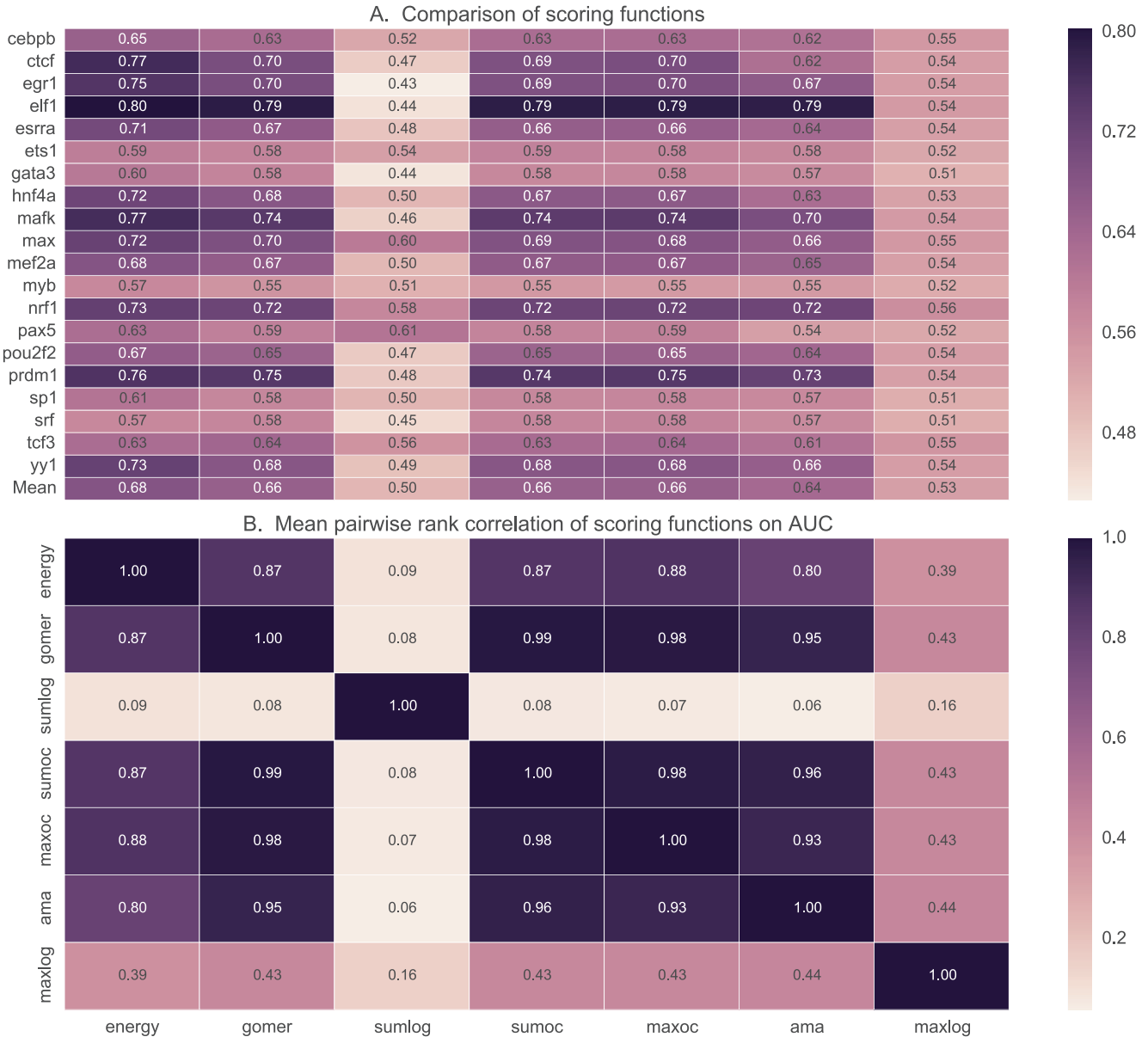

Figure 7. Effect of scoring function on motif ranking using AUC statistic. A. For each TF, the mean AUC score is computed for each of the scoring functions used. In **B**, we show how the ranks assigned to various motifs for a given TF by each scoring function are correlated. It displays the pairwise rank correlation for all TFs in **A**. *Sumlog*: Sum log-odds function, *Sumoc*: sum occupancy score and *Maxoc*: maximum occupancy.

difference (p=0.85, Wilcoxon rank-sum test) between sum occupancy and maximum (Table 2), contrary to a claim by Orenstein *et al.*[28]. When using MNCP, there is a higher rank correlation among the scores assigned by the different scoring functions except log-odds scoring (Figure 8B). When using AUC or MNCP statistic, Ctcf, Egr1 and Hnf4a score significantly higher in energy while other TFs like Pou2f2 and Esrra, the preference is reversed. These observations are reflective of the inherent features of the scoring functions or the statistics used.

Motif length and information content

Motif length has little bearing on the quality of motif, independent of other factors. However, specific motifs with very high IC such as those from POUR have a lower performance (Figure 9). In the same light, those motifs with low IC also have a lower performance *in vivo*.

The heat map in Figure 9 shows how the motif scores from the four discriminative functions correlate with motif length, full-length

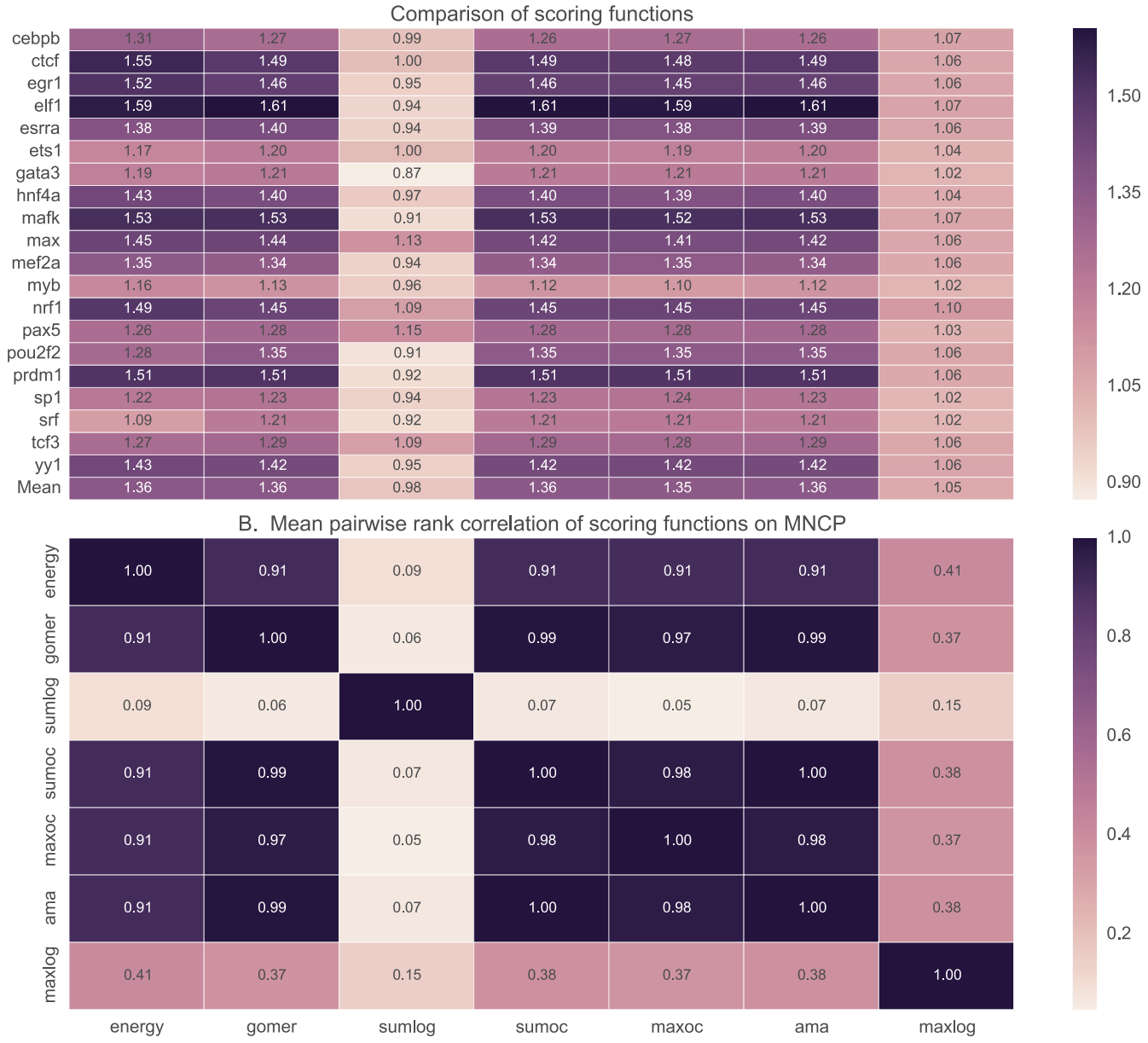

Figure 8. Effect of scoring function on motif ranking based on MNCP statistic. See caption in Figure 7 for details.

Table 2. Mean scores and standard deviation (SD) of AUC and MNCP for scoring functions. For each transcription factor, the median and mean for AUC or MNCP are computed for all the available motifs. *Sumlog*: Sum log-odds function, *Sumoc*: sum occupancy and *Maxoc*: maximum occupancy.

Statistic	Energy	GOMER	Sumlog	Sumoc	Maxoc
Mean AUC	0.68	0.66	0.5	0.66	0.66
Median AUC	0.7	0.67	0.48	0.64	0.64
AUC SD	0.15	0.15	0.11	0.15	0.15
Mean MNCP	1.36	1.36	0.98	1.36	1.35
MNCP SD	0.27	0.32	0.14	0.32	0.31

IC and average IC. The examples have no consistent correlation between the IC and the scores (Figure 9A). However, there is a negative correlation between both the total IC and motif length, and the scores except for sum log-odds scoring, which has no significant correlation (p=0.34, correlation p-value). This shows that motif length, rather than the IC of the motifs, negatively influences the scores assigned by these functions. This is not a general rule. Some TFs exemplify a different scenario. For example, Egr1 (Figure 9B) has a positive correlation between IC and scores and a negative correlation with motif length (except for sum log-odds scoring), showing that it has a highly specific binding site[60]. Mef2a, on the other hand, has a positive correlation between motif length and scores showing preference for longer low information motifs (Figure 9C). This could also reflect variability in binding sites.

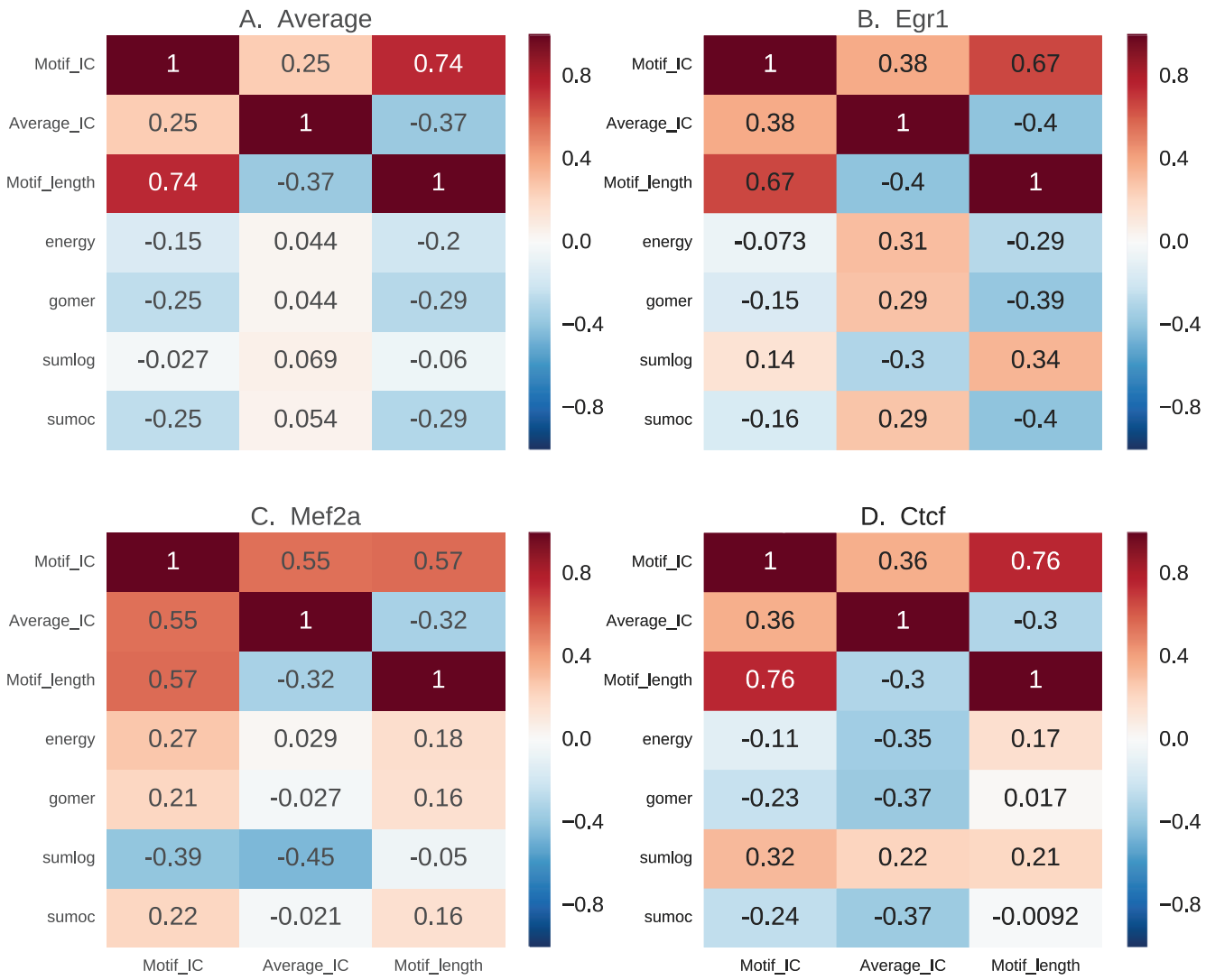

Figure 9. Effect of motif length and IC on scoring functions. In this figure, we show the correlation of motif length, full length information content (IC) and the assessment scores, to determine how performance of scoring functions are influenced by motif characteristics. For each motif, the information content is calculated based on information theory for the whole length and also normalized for length. The results for average motif affinity (AMA) and maximum occupancy are similar to sum occupancy, and are not included.

Ctcf has the highest negative correlation for average IC, with a neutral to positive correlation for motif length (Figure 9D), which may indicate preference for longer low IC motifs.

Comparison of motif databases

We have shown that the effect of scoring algorithms is TF-specific. We also test to see how, overall, the different databases (DBs) are ranked against each other. For TFs with more than one motif in a given DB, we use the best performing one to represent the DB. We also use motif enrichment-based assessment using CentriMo version 4.10.0 to provide more evidence to scoring based techniques' results. Motif enrichment analysis compares how various motifs in foreground sequences are enriched compared with background sequences. In comparing how two or more motifs for the same TF perform, the level of enrichment of the motif in sequences known to contain possible binding sites of the TF compared to some background should provide a measure of the quality of the motif.

Figure 10 provides a summary of ranking of the various databases for the given TFs. We observed that the performance of a majority of the motif databases did not differ much, except for TF2DNA motifs, but the ranking or the performance of individual motifs

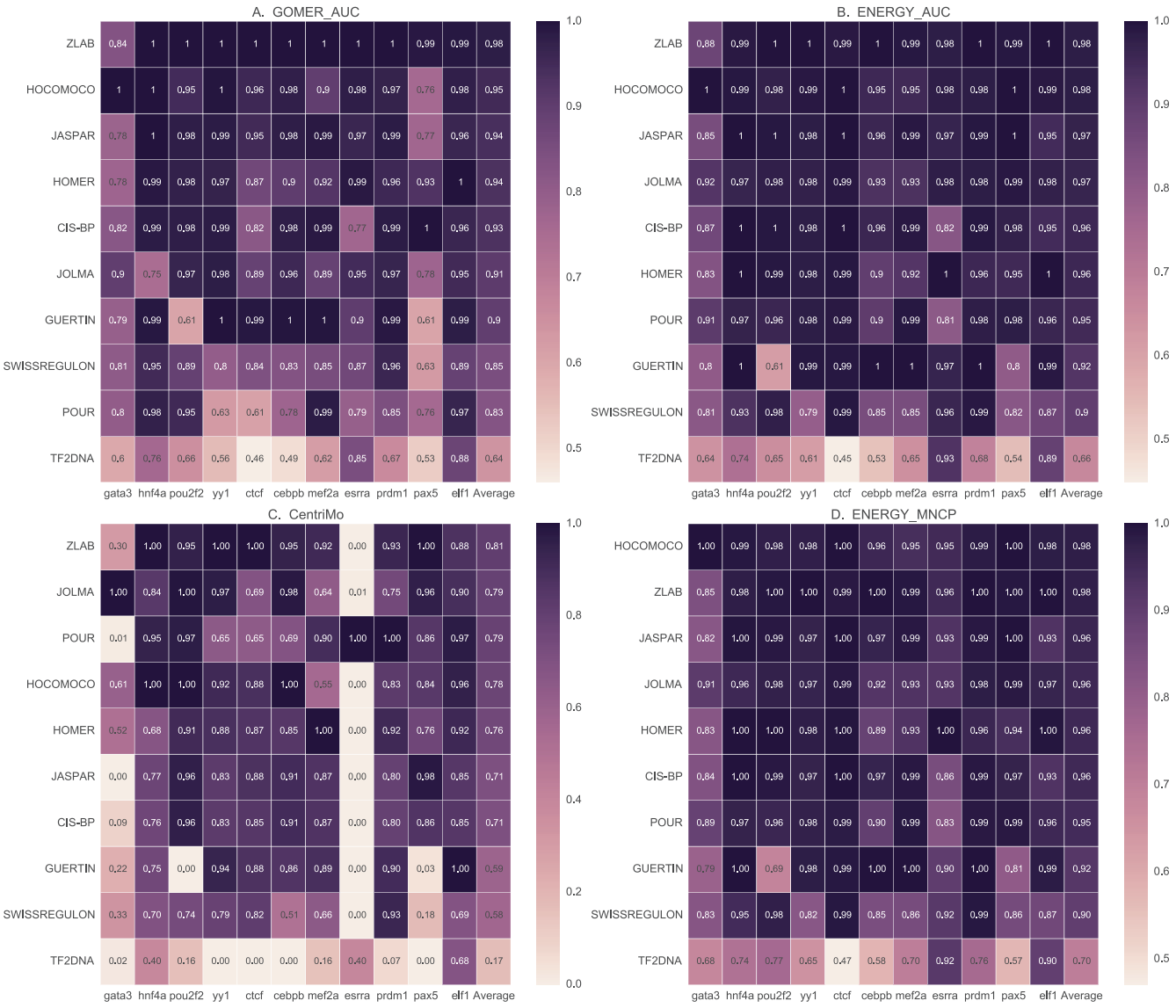

Figure 10. Ranking of motif databases. We compare the motif databases by using the best ranking for each motif using GOMER and energy AUC and MNCP values, and CentriMo enrichment values. For each scoring function, the scores for each TF are normalized by dividing each value with the maximum, which are then averaged to rank the different databases.

differs. This further supports the observation of TF-specific performance of scoring function, algorithms and DBs. It also shows that no single database currently outperforms the others for all TFs. There is agreement in ranking of the best (ZLAB and HOCOMOCO) and worst performing (TF2DNA and SWISSREGULON) DBs. We observe that, compared with GOMER (Figure 10A), the ranks for most DBs remain the same when using energy (Figure 10B) except for POUR and JOLMA. This shows that motifs from these DBs, or at least the best performing ones, may be favoured by energy scoring. It is also noteworthy that POUR and GUERTIN DB motifs, discovered and tested on ENCODE ChIP-seq data, generally perform poorly.

Effect of PBM data on motif assessment

To test whether the conclusions of the paper are only linked to ChIP-seq data, we re-ran the whole analysis using PBM data from the UniPROBE database[13]. It is important to note that we only found 9 TFs that had PBM data from the set used in ChIP-seq analysis.

Since this may bias the comparison, we compared with a similar set in ChIP-seq and found the observations below were not affected by the difference in number of TFs used. These observations include:

1. A much higher energy score in PBM (Figure S1 and Figure S2) compared with ChIP-seq (Figure 7 and Figure 8). We also observe a much lower correlation between the energy and the occupancy scores.

2. A stronger negative correlation between the occupancy scores and motif length -0.47 compared with -0.28 of energy scoring (Figure S3), an observation not made when using ChIP-seq data (Figure 9). This may actually explain observation 1.

3. Motifs generated using the PBM technique perform best when using occupancy scores with MNCP or energy scores with AUC or MNCP, except when occupancy scoring and AUC are used (Figure S4). Poor ranking of UniPROBE PBM-derived motifs by GOMER-AUC may be linked to the fact that they penalize long motifs – UniPROBE motifs are know to be long (mostly over 14bp).

4. Energy scoring with either MNCP or AUC, or occupancy scoring with MNCP display similar behaviour: a preference for specific motifs, which may be longer or have a higher IC. This supported by the high negative correlation between motif length and occupancy scores with AUC (Figure S3).

5. TF2DNA motifs perform better when PBM data is used (Figure S4A) compared with ChIP-seq data (Figure 10A), and especially so when GOMER scoring is used together with AUC statistic. It is not immediately clear what the cause of the difference of performance of TF2DNA motifs in PBM and ChIP-seq data is, but the short length (7bp) of TF2DNA motifs and the fact that they were generated *in vitro* could provide some explanation given that PBM data are generated in 8-mers and PBM is also an *in vitro* technique.

Discussion

We have described a comparative analysis on the effect of scoring functions, ChIP-seq test data processing and statistics on motif assessment. We chose to focus on TF binding models represented as a PWM, since it is most commonly used. The review reveals the complexity of the motif assessment problem, showing no appropriate solution is available so far. The available techniques focus on testing motif algorithms or the experimental techniques used, but little has been done to compare the available motifs for a given TF. There is a need for a tool, accessible and easy to use by end-users, to aid in choosing motifs.

The use of 100 or 250bp sequence length provides necessary discrimination for the TFs tested (Figure 3). The performance was also found to be TF specific, an observation that could reflect inherent binding behaviour; direct, indirect or cooperative binding of the TF. This supports the observation that direct binding can be inferred from ChIP-seq peaks[56]. We also confirm that 100bp provides acceptable specificity in motif assessment given that most TF binding sites are less than 30bp[57].

Since TF binding is cell line specific[61], users should be aware of bias caused by use of one cell line in an assessment. We observe that the use of more than one cell line reduces the bias towards cell line specific motifs (Figure 4).

The MNCP rank-order metric is similar to AUC but derived by plotting true positive hits against all sequences' scores. This places emphasis on true positives and therefore is less affected by false positives. Most of the observations from the PBM-based analysis support the conclusion that energy scoring prefers specific motifs (long or with a high IC). We also observe an agreement when energy scoring is used with AUC and MNCP, or occupancy scoring. In MNCP, the preference for specific motifs is expected because the MNCP score provides a rank-based measure of the ratio of mean occurrence of a motif in test sequences and a random set. These observations are not conclusive and further research may be required. Although there is no clear winner among the scoring function, occupancy-based (GOMER, AMA, sum and max) and energy scoring functions are preferred. We recommend, based on the presented evidence, using occupancy scoring with the MNCP statistic or energy scoring with normal AUC or the MNCP statistic.

There is no significant correlation (p=0.513, correlation p-value) between the IC and the motif scores (Figure 9). This compares with the conflicting observations that the best-quality motifs may have low IC motifs[5], or high IC motifs[62]. Weirauch *et al.* did not normalize for motif length, which results in high IC motifs that are generally longer but not necessarily more specific[5]. A shorter motif with higher IC per position will be more specific but have lower total IC. We argue that the effect of IC on motif quality is dependent on the TF binding behaviour. TFs with short and specific binding sites will favour high IC while those with long and variable binding sites will be more accurately modelled with low IC. Furthermore, it has been shown the low IC flanking motif sites contribute to specificity of binding *in vivo*[62]. We have also shown that the techniques used in motif assessment have a direct effect on motif discovery. We observe how motifs generated from similar data using the same techniques

could have highly variable performance in POUR, ZLAB and GUERTIN motifs (Figure 10). This difference in quality can be explained by motif discovery algorithms used, data processing as well as the assessment techniques used in each motif discovery pipeline. POUR motifs are learned from full-length sequences of the top 250 peaks using five motif finding algorithms (MEME, MDscan, Trawler, AlignAce and Weeder)[41], the ZLAB group used 100bp of the top 500 sequences centred on the ChIP-seq peaks using MEME-ChIP[63], while GUERTIN reports the top 5 motifs for each technique generated using MEME. For quality assessment, POUR[41] used a TFM-PVALUE[64] to match motifs against the testing ChIP-seq data set and the most enriched motifs against a background composed of intergenic non-repetitive regions. ZLAB group used FIMO[55], which uses a log likelihood score for motif scanning.

The worst performing motifs, from TF2DNA, are generated from 3D models of TF from experimental or homology-modelled PDB files. When generating these models, they determined the accuracy of the models based on similarity to UniPROBE and JASPAR motifs at a given threshold. They claimed their technique successfully identifies true motifs 41–81% of the time depending on the quality of templates used in modelling 3D structures. We speculate that part of the reason for this low performance could be use of motif comparison against 'reference motifs' as a measure of motif quality, in addition to being *in vitro* derived. Better performance of TF2DNA motifs in PBM data (Figure S4) further supports this observation. Although JASPAR and UniPROBE are widely used, reliance on reference motifs is problematic, especially with the advent of motif databases like HOCOMOCO and CIS-BP that have motifs with better prediction quality. As a general principle, it is not reasonable to use historical data as a benchmark for assessing potentially better new methods.

We also show that the choice of data used in motif assessment has a direct effect on the ranks of the motifs. It goes without saying that PBM-derived motifs will perform better when tested with PBM data or for ChIP-seq based motifs tested on ChIP-seq data. The main criteria for choosing the test data should be based on the intended use of the motifs. In addition, we confirm the effect of negative (background) sequences in motif assessment, an effect well known in motif discovery.

We have confirmed that motif assessment has transcription-specific variability, an observation previously made[65]. Assessments should be less focused on how a particular motif database or algorithm performs but on how different motifs, for a particular TF, perform compared to the other potential motifs. For the end user, no single database can provide the sole measure of quality of new data. This raises the need for collation of the different motifs tested using a variety of motif assessments to provide information to the end user on their ranks.

Conclusions

We have demonstrated that the scoring techniques used in motif assessment influence ranking of motifs in a TF-specific manner. Motif assessments and tools developed to date have focused on comparing algorithms, experimental techniques or databases. This does not help the user choose which motif to use for a given TF. Some TFs reviewed here have at least 44 PWM motifs available, raising the need for a tool that can be utilized to rank these motifs. We have also shown that data processing as well as motif assessment can have a significant influence on the quality of motifs derived. Therefore, the choice of an assessment approach should be given as much thought as that of the motif discovery algorithm to use. We have also shown the effect of IC on motif quality is influenced by TF binding behaviour.

In short, a single measure of motif quality is likely to remain elusive, pointing to the need for tools and methods for comparative analysis using multiple methods. Lessons learned from this analysis will be useful in a number of ways. Firstly, we are working on a web-based application that can allow users to compare motifs available in different databases for a specific TF. Secondly, we intend to extend the motif by comparison approach to avoid 'reference motifs' bias. Thirdly, we have shown the effect of motif scoring on motif discovery. We intend to use the robust motif assessment techniques we introduce to improve motif finding.

Data and software availability

Data, software, supplementary files and documentation for 'Transcription factor motif quality assessment requires systematic comparative analysis' are available from Github: https://github.com/kipkurui/Kibet-F1000Research.

Archived files at the time of publication are available from Zenodo: doi: 10.5281/zenodo.46440[66]

Author contributions

CK designed and performed the analysis and wrote the first draft. PM supervised the work and contributed to subsequent drafts. All authors read and approved the final manuscript.

Competing interests

No competing interests were disclosed.

Grant information

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.

Supplementary information

This section provides supporting figures for the paper.

Figure S1. Effect of scoring function on motif ranking using AUC statistic for PBM data. A. For each transcription factor (TF), the mean AUC score is used to represent it for each scoring functions used. In **B**, we show how the ranks assigned to various motifs for a given TF by each scoring function are correlated. It displays the pairwise rank correlation for all TFs in **A**. *Sumlog*: Sum log-odds function, *Sumoc*: sum occupancy score and *Maxoc*: maximum occupancy.

Figure S2. Effect of scoring function on motif ranking based on MNCP statistic in PBM data. See caption in Figure S1 for details.

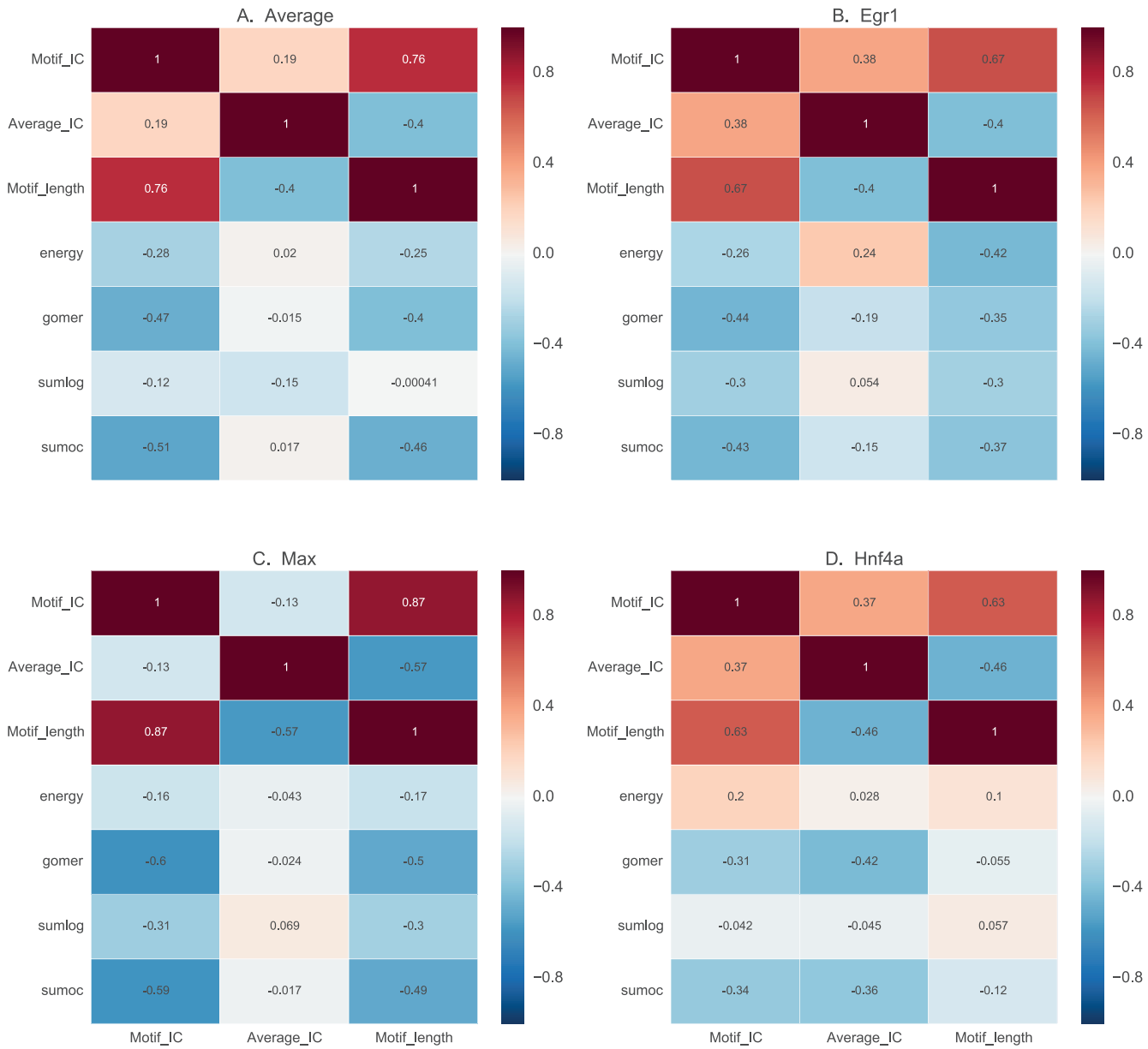

Figure S3. Effect of motif length and IC on scoring functions using PBM data. In this figure, we show the correlation of motif length, full length information content (IC) and the assessment scores, to determine how performance of scoring functions are influenced by motif characteristics. For each motif, the information content is calculated based on information theory for the whole length and also normalized for length. The results for average motif affinity (AMA) and maximum occupancy are similar to sum occupancy, and are not included.

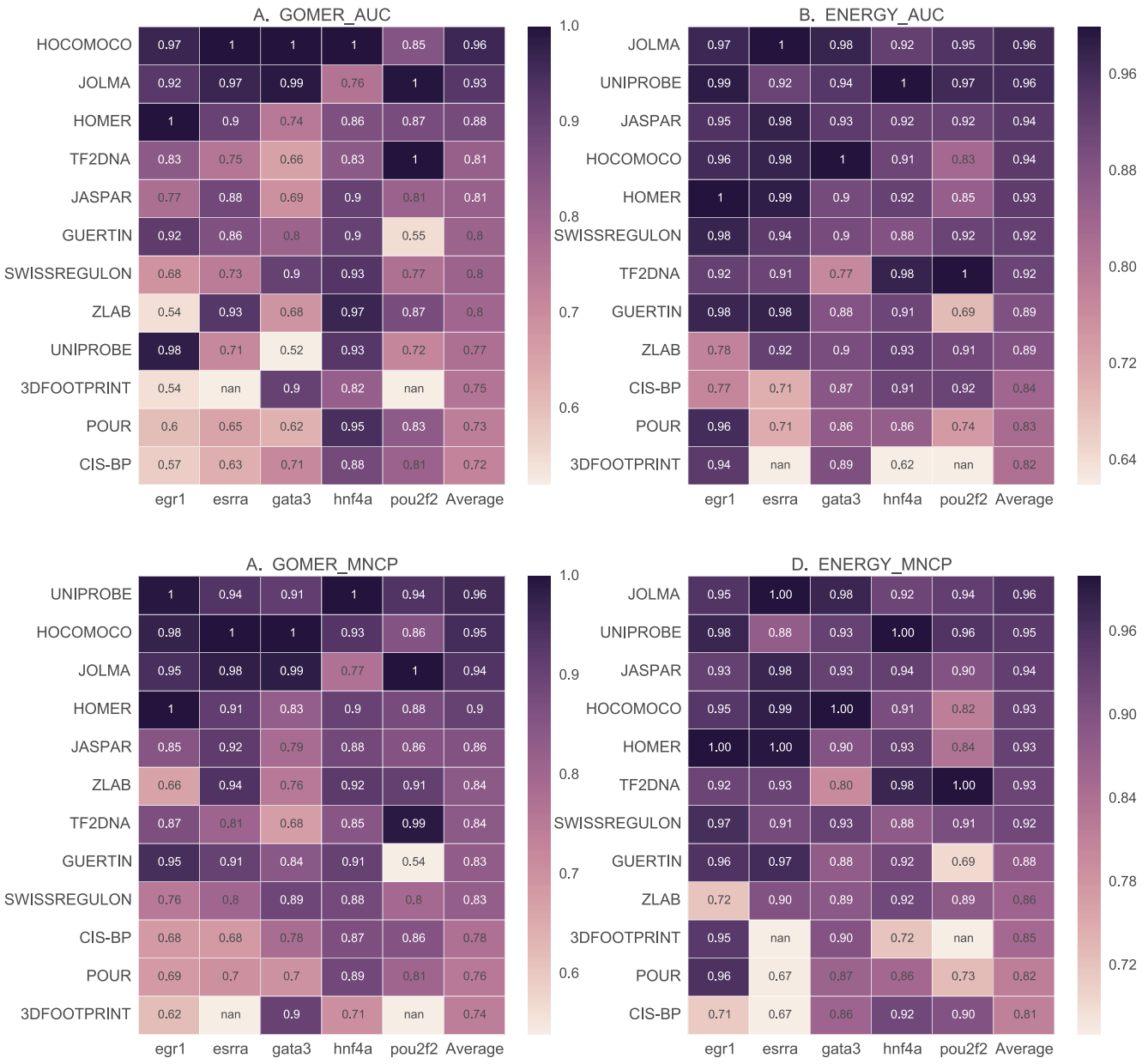

Figure S4. Ranking of motif databases when based on PBM data. We compare the motif databases by using the best ranking for each motif using GOMER and energy AUC and MNCP values, and CentriMo enrichment values. For each scoring function, the scores for each TF are normalized by dividing each value with the maximum, which are then averaged to rank the different databases.

References

1. Annala M, Laurila K, Lähdesmäki H, *et al.*: **A linear model for transcription factor binding affinity prediction in protein binding microarrays.** *PLoS One.* 2011; **6**(5): e20059.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Siddharthan R: **Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix.** *PLoS One.* 2010; **5**(3): e9722.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Wang LS, Jensen ST, Hannenhalli S: **An interaction-dependent model for transcription factor binding.** *Systems Biology and Regulatory Genomics.* 2006; **4023**: 225–234.
Publisher Full Text
4. Zhao Y, Granas D, Stormo GD: **Inferring binding energies from selected binding sites.** *PLoS Comput Biol.* 2009; **5**(12): e1000590.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Weirauch MT, Cote A, Norel R, *et al.*: **Evaluation of methods for modeling transcription factor sequence specificity.** *Nat Biotechnol.* 2013; **31**(2): 126–34.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Mordelet F, Horton J, Hartemink AJ, *et al.*: **Stability selection for regression-based models of transcription factor-DNA binding specificity.** *Bioinformatics.* 2013; **29**(13): i117–i125.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Mathelier A, Wasserman WW: **The next generation of transcription factor binding site prediction.** *PLoS Comput Biol.* 2013; **9**(9): e1003214.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Keilwagen J, Grau J: **Varying levels of complexity in transcription factor binding motifs.** *Nucleic Acids Res.* 2015; **43**(18): e119.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Schneider TD, Stephens RM: **Sequence logos: a new way to display consensus sequences.** *Nucleic Acids Res.* 1990; **18**(20): 6097–6100.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers**. *Proc Int Conf Intell Syst Mol Biol.* AAAI, 1994; **2**: 28–36.
PubMed Abstract
11. Bailey TL: **DREME: motif discovery in transcription factor ChIP-seq data.** *Bioinformatics.* 2011; **27**(12): 1653–1659.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Jin VX, Apostolos J, Nagisetty NS, *et al.*: **W-ChIPMotifs: a web application tool for *de novo* motif discovery from ChIP-based high-throughput data.** *Bioinformatics.* 2009; **25**(23): 3191–3193.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Newburger DE, Bulyk ML: **UniPROBE: an online database of protein binding microarray data on protein-DNA interactions.** *Nucleic Acids Res.* 2009; **37**(Database issue): D77–82.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Jolma A, Yan J, Whitington T, *et al.*: **DNA-binding specificities of human transcription factors.** *Cell.* 2013; **152**(1–2): 327–339.
PubMed Abstract | Publisher Full Text
15. Johnson DS, Mortazavi A, Myers RM, *et al.*: **Genome-wide mapping of *in vivo* protein-DNA interactions.** *Science.* 2007; **316**(5830): 1497–502.
PubMed Abstract | Publisher Full Text
16. Rhee HS, Pugh BF: **Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution.** *Cell.* 2011; **147**(6): 1408–1419.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Zambelli F, Pesole G, Pavesi G: **Motif discovery and transcription factor binding sites before and after the next-generation sequencing era.** *Brief Bioinform.* 2013; **14**(2): 225–37.
PubMed Abstract | Publisher Full Text | Free Full Text
18. Tompa M, Li N, Bailey TL, *et al.*: **Assessing computational tools for the discovery of transcription factor binding sites.** *Nat Biotechnol.* 2005; **23**(1): 137–44.
PubMed Abstract | Publisher Full Text
19. Hu J, Li B, Kihara D: **Limitations and potentials of current motif discovery algorithms.** *Nucleic Acids Res.* 2005; **33**(15): 4899–4913.
PubMed Abstract | Publisher Full Text | Free Full Text
20. Medina-Rivera A, Abreu-Goodger C, Thomas-Chollier M, *et al.*: **Theoretical and empirical quality assessment of transcription factor-binding motifs.** *Nucleic Acids Res.* 2011; **39**(3): 808–824.
PubMed Abstract | Publisher Full Text | Free Full Text
21. Klepper K, Sandve GK, Abul O, *et al.*: **Assessment of composite motif discovery methods.** *BMC Bioinformatics.* 2008; **9**: 123.
PubMed Abstract | Publisher Full Text | Free Full Text
22. Sandve GK, Drabløs F: **A survey of motif discovery methods in an integrated framework.** *Biol Direct.* 2006; **1**: 11.
PubMed Abstract | Publisher Full Text | Free Full Text
23. Sandve GK, Abul O, Walseng V, *et al.*: **Improved benchmarks for computational motif discovery.** *BMC Bioinformatics.* 2007; **8**: 193.
PubMed Abstract | Publisher Full Text | Free Full Text
24. Quest D, Dempsey K, Shafiullah M, *et al.*: **A parallel architecture for regulatory motif algorithm assessment.** *2008 IEEE Int Symp Parallel Distrib Process.* 2008; 1–8.
Publisher Full Text
25. Harbison CT, Gordon DB, Lee TI, *et al.*: **Transcriptional regulatory code of a eukaryotic genome.** *Nature.* 2004; **431**(7004): 99–104.
PubMed Abstract | Publisher Full Text | Free Full Text
26. Zhang Z, Chang CW, Hugo W, *et al.*: **Simultaneously learning DNA motif along with its position and sequence rank preferences through EM algorithm.** *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).* 2012; **7262**: 355–370.
Publisher Full Text
27. Thomas-Chollier M, Herrmann C, Defrance M, *et al.*: **RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets.** *Nucleic Acids Res.* 2012; **40**(4): e31.
PubMed Abstract | Publisher Full Text | Free Full Text
28. Orenstein Y, Linhart C, Shamir R: **Assessment of algorithms for inferring positional weight matrix motifs of transcription factor binding sites using protein binding microarray data.** *PLoS One.* 2012; **7**(9): e46145.
PubMed Abstract | Publisher Full Text | Free Full Text
29. Orenstein Y, Shamir R: **A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data.** *Nucleic Acids Res.* 2014; **42**(8): e63.
PubMed Abstract | Publisher Full Text | Free Full Text
30. Zhong S, He X, Bar-Joseph Z: **Predicting tissue specific transcription factor binding sites.** *BMC Genomics.* 2013; **14**: 796.
PubMed Abstract | Publisher Full Text | Free Full Text
31. Agius P, Arvey A, Chang W, *et al.*: **High resolution models of transcription factor-DNA affinities improve *in vitro* and *in vivo* binding predictions.** *PLoS Comput Biol.* 2010; **6**(9): pii: e1000916.
PubMed Abstract | Publisher Full Text | Free Full Text
32. Slattery M, Zhou T, Yang L, *et al.*: **Absence of a simple code: how transcription factors read the genome.** *Trends Biochem Sci.* 2014; **39**(9): 381–399.
PubMed Abstract | Publisher Full Text | Free Full Text
33. Feingold EA, Good PJ, Guyer MS, *et al.*: **The ENCODE (ENCyclopedia of DNA elements) project.** *Science.* 2004; **9305**(301).
34. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics.* 2010; **26**(6): 841–2.
PubMed Abstract | Publisher Full Text | Free Full Text
35. Chen X, Xu H, Yuan P, *et al.*: **Integration of external signaling pathways with the core transcriptional network in embryonic stem cells.** *Cell.* 2008; **133**(6): 1106–17.
PubMed Abstract | Publisher Full Text
36. Wang J, Zhuang J, Iyer S, *et al.*: **Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors.** *Genome Res.* 2012; **22**(9): 1798–1812.
PubMed Abstract | Publisher Full Text | Free Full Text
37. Chen X, Hughes TR, Morris Q: **RankMotif++: a motif-search algorithm that accounts for relative ranks of K-mers in binding transcription factors.** *Bioinformatics.* 2007; **23**(13): i72–9.
PubMed Abstract | Publisher Full Text
38. Badis G, Berger MF, Philippakis AA, *et al.*: **Diversity and complexity in DNA recognition by transcription factors.** *Science.* 2009; **324**(5935): 1720–1723.
PubMed Abstract | Publisher Full Text | Free Full Text
39. Mathelier A, Zhao X, Zhang AW, *et al.*: **JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles.** *Nucleic Acids Res.* 2014; **42**(Database issue): D142–7.
PubMed Abstract | Publisher Full Text | Free Full Text
40. Zhao Y, Stormo GD: **Quantitative analysis demonstrates most transcription factors require only simple models of specificity.** *Nat Biotechnol.* 2011; **29**(6): 480–483.
PubMed Abstract | Publisher Full Text | Free Full Text
41. Kheradpour P, Kellis M: **Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments.** *Nucleic Acids Res.* 2014; **42**(5): 2976–87.
PubMed Abstract | Publisher Full Text | Free Full Text
42. Kulakovskiy IV, Medvedeva YA, Schaefer U, *et al.*: **HOCOMOCO: a comprehensive collection of human transcription factor binding sites models.** *Nucleic Acids Res.* 2013; **41**(Database issue): D195–202.
PubMed Abstract | Publisher Full Text | Free Full Text
43. Pachkov M, Erb I, Molina N, *et al.*: **SwissRegulon: a database of genome-wide annotations of regulatory sites.** *Nucleic Acids Res.* 2007; **35**(Database issue): D127–D131.
PubMed Abstract | Publisher Full Text | Free Full Text
44. Pujato M, Kieken F, Skiles AA, *et al.*: **Prediction of DNA binding motifs from 3D models of transcription factors; identifying TLX3 regulated genes.** *Nucleic Acids Res.* 2014; **42**(22): 13500–12.
PubMed Abstract | Publisher Full Text | Free Full Text
45. Heinz S, Benner C, Spann N, *et al.*: **Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities.** *Mol Cell.* 2010; **38**(4): 576–589.
PubMed Abstract | Publisher Full Text | Free Full Text

46. Contreras-Moreira B: **3D-footprint: a database for the structural analysis of protein-DNA complexes.** *Nucleic Acids Res.* 2010; **38**(Database issue): D91–D97.
PubMed Abstract | Publisher Full Text | Free Full Text

47. Guertin MJ, Martins AL, Siepel A, *et al.*: **Accurate prediction of inducible transcription factor binding intensities *in vivo*.** *PLoS Genet.* 2012; **8**(3): e1002610.
PubMed Abstract | Publisher Full Text | Free Full Text

48. Weirauch MT, Yang A, Albu M, *et al.*: **Determination and inference of eukaryotic transcription factor sequence specificity.** *Cell.* 2014; **158**(6): 1431–1443.
PubMed Abstract | Publisher Full Text | Free Full Text

49. Granek JA, Clarke ND: **Explicit equilibrium modeling of transcription-factor binding and gene regulation.** *Genome Biol.* 2005; **6**(10): R87.
PubMed Abstract | Publisher Full Text | Free Full Text

50. Foat BC, Morozov AV, Bussemaker HJ: **Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE.** *Bioinformatics.* 2006; **22**(14): e141–9.
PubMed Abstract | Publisher Full Text

51. Bailey TL, Boden M, Buske FA, *et al.*: **MEME SUITE: tools for motif discovery and searching.** *Nucleic Acids Res.* 2009; **37**(Web Server issue): W202–W208.
PubMed Abstract | Publisher Full Text | Free Full Text

52. Clarke ND, Granek JA: **Rank order metrics for quantifying the association of sequence features with gene regulation.** *Bioinformatics.* 2003; **19**(2): 212–218.
PubMed Abstract | Publisher Full Text

53. van Heeringen SJ, Veenstra GJ: **GimmeMotifs: a *de novo* motif prediction pipeline for ChIP-sequencing experiments.** *Bioinformatics.* 2011; **27**(2): 270–271.
PubMed Abstract | Publisher Full Text | Free Full Text

54. Lesluyes T, Johnson J, Machanick P, *et al.*: **Differential motif enrichment analysis of paired ChIP-seq experiments.** *BMC Genomics.* 2014; **15**(1): 752.
PubMed Abstract | Publisher Full Text | Free Full Text

55. Grant CE, Bailey TL, Noble WS: **FIMO: scanning for occurrences of a given motif.** *Bioinformatics.* 2011; **27**(7): 1017–8.
PubMed Abstract | Publisher Full Text | Free Full Text

56. Bailey TL, Machanick P: **Inferring direct DNA binding from ChIP-seq.** *Nucleic Acids Res.* 2012; **40**(17): e128.
PubMed Abstract | Publisher Full Text | Free Full Text

57. Wilbanks EG, Facciotti MT: **Evaluation of algorithm performance in ChIP-seq peak detection.** *PLoS One.* 2010; **5**(7): e11471.
PubMed Abstract | Publisher Full Text | Free Full Text

58. Takahashi K, Hayashi N, Shimokawa T, *et al.*: **Cooperative regulation of Fc receptor gamma-chain gene expression by multiple transcription factors, including Sp1, GABP, and Elf-1.** *J Biol Chem.* 2008; **283**(22): 15134–41.
PubMed Abstract | Publisher Full Text | Free Full Text

59. Bengtsen M, Klepper K, Gundersen S, *et al.*: **c-Myb Binding Sites in Haematopoietic Chromatin Landscapes.** *PLoS One.* 2015; **10**(7): e0133280.
PubMed Abstract | Publisher Full Text | Free Full Text

60. Kubosaki A, Tomaru Y, Tagami M, *et al.*: **Genome-wide investigation of *in vivo* EGR-1 binding sites in monocytic differentiation.** *Genome Biol.* 2009; **10**(4): R41.
PubMed Abstract | Publisher Full Text | Free Full Text

61. Lower KM, De Gobbi M, Hughes JR, *et al.*: **Analysis of sequence variation underlying tissue-specific transcription factor binding and gene expression.** *Hum Mutat.* 2013; **34**(8): 1140–1148.
PubMed Abstract | Publisher Full Text

62. Orenstein Y, Mick E, Shamir R: **RAP: accurate and fast motif finding based on protein-binding microarray data.** *J Comput Biol.* 2013; **20**(5): 375–82.
PubMed Abstract | Publisher Full Text | Free Full Text

63. Machanick P, Bailey TL: **MEME-ChIP: motif analysis of large DNA datasets.** *Bioinformatics.* 2011; **27**(12): 1696–1697.
PubMed Abstract | Publisher Full Text | Free Full Text

64. Touzet H, Varré JS: **Efficient and accurate P-value computation for Position Weight Matrices.** *Algorithms Mol Biol.* 2007; **2**: 15.
PubMed Abstract | Publisher Full Text | Free Full Text

65. Zhang Y, He Y, Zheng G, *et al.*: **MOST+: A *de novo* motif finding approach combining genomic sequence and heterogeneous genome-wide signatures.** *BMC Genomics.* 2015; **16**(Suppl 7): S13.
PubMed Abstract | Publisher Full Text | Free Full Text

66. Kibet CK: **Kibet-F1000Research: Kibet-F1000Research V2.0.** *Zenodo.* 2016.
Data Source

Open Peer Review

Current Referee Status:  

Version 2

Referee Report 01 April 2016

doi:10.5256/f1000research.8713.r13169

Trevor W. Siggers
Department of Biology, Boston University, Boston, MA, USA

I accept this revised version of the manuscript. I am satisfied with the updates provided by the authors.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 31 March 2016

doi:10.5256/f1000research.8713.r12883

Jan Grau
Institute of Computer Science, Martin Luther University of Halle-Wittenberg, Halle, Germany

I am happy to state that the authors thoroughly addressed most of my concerns regarding the previous version of the manuscript.
However, I still have a few minor comments, which the authors might consider, as listed in the following.

1. In the introduction, the authors state that Chip-seq has improved "resolution to single-nucleotide level", whereas at the beginning of the Results section, they state that "a successful ChIP-seq experiment localizes binding within about 100bp of the true site", which still is quite accurate but not exactly single-nucleotide accuracy. Please clarify.

2. In section "Background", sub-section "Motif comparison", the authors list as possible measures of motif divergence "sum of square deviation, Euclidean distance". As the former is just the square of the latter, I would suggest to give another example (for instance, correlation-based meaures) in the introduction of that sub-section.

3. In section "Methods", sub-section "Data", the authors now describe the selection of negative sequences, which were "extracted 500bp downstream from the highest coordinate".

   a) However, as ChIP-seq peaks lack an orientation, I wonder which direction "downstream" refers to. Does this mean that the authors considered the forward strand of the available genomic sequence, or do they also take, e.g., the orientation of closely located genes into account?

b) In the response, the authors explain that "our focus was to get negative sequences [...] which maintains the nucleotide composition". As I mentioned in my previous review, sequences located 500bp downstream of a transcription factor binding site might already be located in coding sequence, which affects nucleotide composition. Could the authors please comment on this issue?

4. In section "Methods", sub-section "Data", the authors explain that they "only found nine TFs that had comparable data in ChIP-seq and PWM". However, there are additional, similar data sets for at least Foxo1, Zfx, Tbx5, and Nr5a2 available (see Grau *et al*., 2013 for details). Please clarify.

5. In equation 1, parentheses might help to spot the argument of the product.

6. In equation 4, the parameter μ is not explained.

7. In Figure 3, the authors discuss the influence of sequence length on AUC. However, I would consider the influence of sequence length on the ranking of motifs more relevant for the topic of the manuscript.

8. In section "Results", sub-section "Effect of PBM data on motif assessment", 5th item of the list, the authors state that "PBM data are generated in 8-mers". As far as I know, PBMs are typically designed such that they cover all 10-mers (see, e.g., http://the_brain.bwh.harvard.edu/pbm.html). In addition, the authors might clarify which data they used for the assessment (the probe sequences and intensities, I assume?).

References

1. Grau J, Posch S, Grosse I, Keilwagen J: A general approach for discriminative de novo motif discovery from high-throughput data.*Nucleic Acids Res*. 2013; **41** (21): e197 PubMed Abstract | Publisher Full Text

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Version 1

Referee Report 15 January 2016

doi:10.5256/f1000research.7983.r11721

Jan Grau
Institute of Computer Science, Martin Luther University of Halle-Wittenberg, Halle, Germany

The manuscript "Transcription factor motif quality assessment requires systematic comparative analysis" by Kibet and Machanick addresses the assessment of transcription factor binding motifs. This question is especially important for selecting appropriate motifs for computational predictions given the large number of different motifs for the same transcription factor available from databases. Kibet and Machanick specifically consider different measures for motif scoring and assessment, and investigate different

factors that might influence the assessment and, hence, the chosen motif.

The topic is of great relevance in any research dealing with sequence motifs and a systematic analysis of the factors influencing their assessment may help to develop a standardized framework for motif assessment.
However, I have several reservations regarding the current version of the manuscript as outlined below.

As a general comment (that does not necessarily require a response by the authors), I found it slightly disappointing that the present manuscript does pose many important questions and potential obstacles in motif assessment, but does not provide a solution, be it guidelines for reasonable motif assessment or be it even a platform for performing such analyses.

DATA:

1. I wonder why the authors decided to only consider ChIP-seq data but no *in-vitro* data (PBMs or SELEX). While *in-vivo* binding may of greater relevance for many applications, problems like cell type-specificity of motifs (also addressed by the authors) would have a minor influence. In addition, competitive or interaction effects with other transcription factors might be ruled out. Finally, some of the motif sources considered derive their PWMs from *in-vitro* data. In summary, an analysis also using *in-vitro* data might affect the conclusions of the paper.

2. For all ChIP-seq data sets under consideration, the authors extract (only) the top 500 highest scoring sequences for the assessment. This may have largely differing effects for different transcription factors, where, for instance, one transcription factor might have several hundred ChIP-seq positive regions, whereas another transcription factor might have tens of thousands of ChIP-seq positive regions. Hence, in one case also lowly occupied sequences are collected whereas in the other case, the positive data set may only contain the most stringent binding sequences. This may affect all downstream analyses and, for instance, could be one of the reasons why the authors observe transcription factor-specific effects for some factors. Hence, I would strongly suggest to conduct the analysis with a transcription factor-specific selection of sequences (where the simplest idea might be to use just a percentile).

3. a) The authors state that they construct a "similar" negative set. Here, the authors should clearly define what "similar" means, how sequences are selected, and how many negative sequences are in the set.
b) In addition, the specific selection of negative sequences described by the authors (500bp "downstream", where "downstream" is also unclear as ChIP-seq regions lack an orientation), might introduce a specific bias, because under the assumption that transcription factor binding sites are often located close to the transcription start, which might mean that the negative sequences may already be coding and, hence, per se different from promoter sequences.
c) Finally, from my experience, the choice of the negative data set strongly affects the performance assessment of motifs. Hence, the authors might consider to test an additional set of negative sequences (e.g., di-nucleotide shuffled positive sequences) in their analysis.

METHODS:

4. The "Methods" section, especially formulas needs substantial revision:
a) In general, notation should be harmonized between the different formulas. For instance, the sequence S appears with different indexes with different meanings; the indicator function is denoted by S_i(b,m) in eqn. (5) and by I(S_{i,b}) in equation (6).

In addition:
b) In eqn. (1), parentheses are missing around (1 - P(...)). In addition, the notation S_{t+1:t+k}^{i} is not explained
c) In eqn. (2), it is unclear if i and [S_{t=i}] are indexes or if this should denote a product of \theta, i and S_{t=i} (which I consider unlikely). In addition, the variable t (in the index) is neither bound nor explained.
d) In eqn. (3), the upper limit of the sum is |s|, where it should be |S|, I assume. In addition, there seems to be something missing (a \theta?) in the product.
e) Before eqn. (5), the authors refer to T as the length of the sequence. However, considering the formula, the length should be L, and the first sum from A to T refers to the alphabet. In addition, eqn. (6) again denotes the sum over the alphabet differently.
f) In eqn. (5), the text refers to sequence S but the formula to sequence S_i
g) In eq. (6), the variable P_b is not defined (the authors later only refer to p, which might have the same meaning). In addition, the authors to not explain, which background distribution they use in the assessment, which will be relevant, e.g., for the results presented in Fig. 6.

5. The energy scoring framework (eqn. 4 and 5) and the LogOdds scoring framework are formally defined only for sub-sequences and it remains unclear how these are applied to longer sequences from ChIP-seq. Are those subjected to the occupancy definitions (maximum and/or average) as well?

6. LogOdds scoring is referred to as "Log likelihood scoring" in the section's title (page 6, left column), which is not fully correct.

7. On page 6, right column, second paragraph, the authors state that they "wish to check the usefulness of correlation in motif assessment" (which I would find interesting), but I did not find any results regarding correlation as performance measure in the results.

RESULTS:

8. In several cases, the figure captions are too minimalistic to understand the contents of the figure. I would suggest to spend a few more sentences in the captions to explain the main idea of each figure. In addition, not all of the abbreviations are explained in the caption of Fig. 6.

9. On page 6, penultimate paragraph, the authors state that "the Foxa motif from the POUR data set is significantly differentially enriched only in the A549 cell line", which I could not read from Fig. 4. Please clarify.

10. On page 8, right column, the authors state that "MNCP prefers specific motifs, which will have more true positives". Could the authors elaborate on these findings and also possibly give an (mathematical) explanation?

11. In Fig. 6, the authors show AUC values for different motifs and scoring functions.
a) First, it remains unclear which data sets have been used in this analysis for the different transcription factors. Is it just the average over all motifs and data sets for each factor?
b) Second, I did, unfortunately, not get the general idea of this analysis. If I understood it correctly, the main question of this manuscript is to study the effects of different factors on motif assessment with the goal of selecting the most appropriate motif for a given transcription factor. However, here it seems to be that exactly this information is averaged out. Wouldn't be the more interesting question how the scoring functions affect the ranking (by AUC) of the different motifs for each transcription factor?

12. On page 10, left column, the authors state that they "did not observe any significant difference (p=0.85, Wilcoxon rank-sum test) between sum occupancy and maximum (Table 3)". However, I did not find maximum occupancy listed in Tab. 3.

13. On page 11, right column, the authors state that Egr1 has strong positive correlation between IC and scores. However, I found this correlation not too strong for Average_IC and in most cases not even positive for Motif_IC.

14. a) In remains unclear, what exactly is shown in Fig. 8. I speculate that the authors computed the correlation of AUC values, IC and motif length for different data sets and motifs? Or is it really correlation between occupancy/energy and IC/length?
b) In addition, most of the entries of the heatmaps show correlations between the occupancies/energy, which, however, is not discussed. If correlation between occupancies/energy is not of interest, the authors might consider omitting all but the first three rows of the heatmaps.
c) Further, I wonder why the correlation between identical entries (e.g., Motif_IC with Motif_IC) is not equal to 1 in panel A.

15. On page 12, second paragraph, the authors explain that they used the best performing motif to represent each database. However, this will introduce a bias towards larger databases, because these may contain a larger number of motifs for a transcription factor and, hence, are allowed to try a larger number of options, of which the best is chosen. I would suggest to use another, less biased statistic (e.g., the median) instead/in addition.

16. The authors also use CentriMo scoring for comparing databases, which they did not consider before, and I wonder what is the reasoning behind using CentriMo in this case (and not before).

17. In Figure 9, panel C, the authors rank the databases by average CentriMo score, while the magnitude of scores differs greatly between transcription factors and, hence, is dominated by data sets with large scores (e.g., cebp). I would suggest to level the influence of transcription factors, for instance by dividing the values in each column by their maximum value before averaging.

18. On page 12/14, the author state that "This supports our view that use of motif comparison against ‘reference motifs’ as a measure of motif quality is not reliable". While I agree with the general conclusion of the authors, I do not see why the performance of TF2DNA supports this conclusion. If only 41-81% of the TF2DNA motifs are correct (according to comparison against reference motifs), I would have expected a lower performance compared to the other databases.

OTHER/MINOR:

19. In section "Background", second paragraph, the authors refer to Weirauch *et al*., stating that a well-trained PWM performs comparably to more complex models. While this correctly describes the finding of Weirauch *et al*., several publication in the meantime came to different conclusions (e.g., Kulakovskiy *et al*., 2013[1]; Mathelier & Wasserman, 2013[2]; Mordelet *et al*., 2013[3]; Keilwagen & Grau, 2015[4]). Hence, the authors might consider to make this statement more balanced.

20. In section "Background", fourth paragraph, the authors state that "the quality of models derived has not improved in a comparative manner". I am not fully sure if I understand the statement correctly, but if

the authors mean that the experimental techniques have improved, but the motifs did not (or much less), I would challenge this statement and at least encourage the authors to provide a reference.

21. The authors should provide a list (or a link to a list in their repository) of the specific ENCODE data sets used in the analysis.

22. Table 1: Chen2008 should be ChIP-seq data.

23. As performance measures, the authors consider the area under the ROC curve and MNCP. While the former might be familiar for most working in the field, the authors might consider to give a short formal definition of MNCP. In addition, the area under the precision-recall curve might be another useful measure for imbalanced data sets. [However, depending on the construction of the negative data set, the test data might even be balanced.]

24. Typos & Grammar:
- Page 4, second paragraph: "Sandev" should be "Sandve"
- Page 4, 5th paragraph: "Sandelinâ-Wasserman" should be "Sandelin-Wasserman"

References

1. Kulakovskiy I, Levitsky V, Oshchepkov D, Bryzgalov L, Vorontsov I, Makeev V: From binding motifs in ChIP-Seq data to improved models of transcription factor binding sites.*J Bioinform Comput Biol*. 2013; **11** (1): 1340004 PubMed Abstract | Publisher Full Text
2. Mathelier A, Wasserman WW: The next generation of transcription factor binding site prediction.*PLoS Comput Biol*. 2013; **9** (9): e1003214 PubMed Abstract | Publisher Full Text
3. Mordelet F, Horton J, Hartemink AJ, Engelhardt BE, Gordân R: Stability selection for regression-based models of transcription factor-DNA binding specificity.*Bioinformatics*. 2013; **29** (13): i117-25 PubMed Abstract | Publisher Full Text
4. Keilwagen J, Grau J: Varying levels of complexity in transcription factor binding motifs.*Nucleic Acids Res*. 2015; **43** (18): e119 PubMed Abstract | Publisher Full Text

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Author Response 10 Feb 2016

Caleb Kipkurui, Rhodes University, South Africa

Thank you very much for your insightful comments and recommendations. They have helped us improve the paper.

The main aim of this paper is to identify the weaknesses and potential pitfalls in the current techniques used in motif assessment. As part of our conclusions, we state that we will use the findings of this paper to develop a motif assessment platform to address the questions and the gaps. That work is almost done and should be available by March 2016, and is therefore out of the scope of this paper.

1. We focused on ChIP-seq data in assessment since we believe that, for most cases, the final

utility of the motifs learned is predicting *in vivo* binding of the motifs. That said, we agree that data used in testing the motifs does have an effect on the ranking of the motifs. This is an observation we confirm with our re-analysis using PBM data. We have included a section on how assessment in PBM and ChIP-seq are influenced differently (Effect of PBM data on motif assessment).

2. Our choice for top 500 sequences was informed by our understanding of previous research. However we did not make it clear that such prior work supports this point, and we have now cited a reference. As advised, we decided to test if this would affect the results from this analysis. The ChIP-seq peaks we use have a median of 14000 peaks, the highest having 92,258 peaks and a minimum of 101 peaks. Where the number of the available peaks was less than 500, we used all the peaks. Given the median number of peaks, we found 5% of the peaks to be appropriate and we used this when 5% the of the total was more than 500, else we used top 500 peaks (or all of them, for data sets smaller than this). We also tested for 10% of the peaks. In all this, we found that the size of the peaks used had no significant effect on the results obtained. We, therefore, eliminate that as being one of the reasons for cell line specific binding. This may, however, have an effect on cell line specific ranking behaviour even if we did not observe that in our examples, given that the number of peaks differs for a given TF in different cell lines. We will definitely consider this suggestion when developing our tool to avoid any potential bias caused by this.

3. a) The manuscript has been updated. By similar we mean in size and sequence length.

b) In our analysis, downstream is based on the coordinates of the peaks. We extracted sequences located 500bp from the highest coordinate (highest coordinate + 500). Our focus was to get negative sequences which are not expected to contain binding sites for the given TF but which maintains the nucleotide composition. That distance, whether it falls in a promoter region or not, should be appropriate in for our specific analysis.

c) On other negative sequences that can be chosen, we agree that this can have some influence in the analysis. The scores obtained when a negative set generated using dinucleotide shuffled positive sequences were always lower than those from downstream sequences. However, the ranking of the motifs did not change in any significant manner. We expect random negative sets to have a significant influence on the ranks of the motifs and their probable difference in GC content from the binding region makes their appropriateness questionable.

4. The notations used in the formulas have been updated for uniformity. Thank you!

5. For energy scoring, the subsequence with the lowest energy is used to represent the sequence while for logOdds scoring, the score can either be obtained by getting the sum or the maximum score for all the sub-sequences. Clarified in the paper, thank you.

6. Corrections have been made in response to 6, 8 and 9.

7. We have updated the Figure 5 to include details on the usefulness of correlation statistics. We find them to produce significantly different ranks from MNCP or AUC or even between Pearson and Spearman correlations.

8. Done

9. Done

10. We have updated the paper in to add a line giving more information about MNCP. Simply put, the MNCP is a rank-based statistic that determines if the mean occurrence of a motif in test sequences is higher that the mean occurrence in a random set. Each set of sequences is ranked based on the mean occurrence, and the MNCP is calculated by finding the mean of the normalized ratio of the two ranks.

11. We have updated Figure 6 (now Figure 7) to address the comments. Our earlier figure actually averaged the information on the effect of scoring functions on the ranks of the motifs. We have updated by using the rank correlation of the motifs for various TFs to show how it affects ranking.

12. Table 3 (now Table 2) has been updated to include maximum occupancy.

13. On Egr1 motifs correlation of motif IC and scores, we have updated the statement to be in accordance with the data.

14. In Figure 8 (now Figure 9) we have updated the figure to only retain relevant columns. We have also corrected the error that led to identical entries' correlation being more than 1.

15. On why we chose to use the best motif's score to represent a database, we argue that since the focus of this analysis is to test our ability to choose for the best motif, irrespective of the database, we find using the best motif score to represent the DB to be sufficient. Besides, using median will still lead to biased results since DBs with many motifs of low quality and a few of high quality will be poorly ranked.

16. We only introduce CentriMo at a later stage of our analysis as an alternative method of scoring techniques to motif assessment. The focus of the paper was to systematically assess the factors that do influence motif assessment, so we wanted to maintain that focus.

17. We have taken your suggestion on Figure 9 (now 10) to normalize the scores. Thank you.

18. On the performance of TF2DNA, we agree that the low performance would be expected. We also believe that a different approach to motif assessment during motif discovery may have produced better motifs. In addition, testing using PBM data produced a much better performance. This may be a consequence of the motifs being short and only generated using *in vitro* methods.

19. The background section has been updated to include to making the observations balanced and including recent citations.

20. We accept that our statement on the lack of significant improvement of the motifs may have been misleading and unsupported. We have updated it to reflect current evidence.

21. A list of the ENCODE data we use has been added to the repository

22. The source of Chen2008 updated to ChIP-seq from PBM in table 1. Thank you

23. A definition of MNCP has been added to the paper. We had previously tested area under a precision-recall curve and found it to produce similar results to AUC.

24. Typos corrected

Once again, thank you.

Competing Interests: No competing interests were disclosed.

Referee Report 05 January 2016

doi:10.5256/f1000research.7983.r11604

Trevor W. Siggers
Department of Biology, Boston University, Boston, MA, USA

The manuscript by Kibet *et al.* "Transcription factor motif quality assessment requires systematic comparative analysis" addresses an important issue in the field of regulatory genomics, namely how we analyze motif enrichment in genomic datasets. The authors have addressed this issue in a systematic way by compiling many datasets and versions of motifs, and analyzing the impact of different scoring methods.

This type of meta-analysis will be of interest to a wide audience. However, the current manuscript needs considerable revision. In particular, the connections between the data presented and the conclusions reached need to be strengthened and clarified. Furthermore, a lot more clarification about what is being shown in the figures is needed to properly evaluate the conclusions. Below I have outlined specific examples through to Figure 8. The figure legends could definitely use more detail to help clarify what is being shown, and there needs to be more explicit and careful connection between the data and the conclusions (see examples below). I think that the type of analyses contained in this manuscript will be of interest to a wide audience; however, the manuscript needs to be substantially revised.

Table 1. Is Chen2008 databases, Reference 39, really PBM data?

Methods/Data. For each peak file, the 500 highest scored sequences were identified "after eliminating repeat masked sites". It is a little unclear what this statement means. Does that mean that no peak was selected if there was any repeat masked seqeuence within the 50, 100 and 250bp windows? Or was the repeat masked sequence just masked and the genomic window extended to attain the 50, 100 and 250 bp cutoffs? Also, for the negative set, does 'similar' mean length-matched? It was exactly clear how this negative set was constructed.

Figure 3 /results. It was not clear why only a subset of 15 of the Encode ChIP-seq datasets were used and shown here, and how many datasets were used in the 'Average'? Also, the figure caption notes that 'all the motifs for the 15 TFs' were used, but it's not clear how many that was and whether the reported AUC values were averages over their individual AUC values? I little more clarification would be helpful.

Page 6. The authors write, "Unless the interest is tissue-specific binding, if more than one set of data is available, an average should be used". Used for what? For motif discovery?

Figure 4. Why was 'energy scoring' used for this enrichment analysis, while GOMER scoring was used in

Figure 3? Are the results dependent on these scoring differences? If not, then for consistency sake, it would be helpful to limit the enrichment analyses to a single scoring scheme.

Page 6/Figure 4. The authors conclude, “the Foxa motif from the POUR data set is significantly differentially enriched only in the A549 cell line and not so much in the other cell lines”. I have no idea to what the authors are referring here, and this is the only conclusion from Figure 4. There are 5 different FOXA_discX.POUR motifs, all of which seem to score about the same on the different ChIP datasets. There is a FOXA1_2.GUERTIN that seems to be quite different, but this seems like an outlier within the dataset. I do not see how the data supports the contention that there are specific FOXA motifs that are better suited to particular ChIP datasets, it seems that for the most part they agree. Much more clarity is needed here.

Page 8 / Figure 5.
“However, in some situations like Hnf4a and Ctcf, they are not (Figure 5)”. I only see Ctcf data represented in Figure 5, this should be clarified.

“The motifs ranked higher only by MNCP are generally long or with high IC (Table 2)”. It would be much easier to see this if they were indicated somehow in Figure 5, perhaps with arrows are stars or something. Second, these conclusions don’t seem to follow from the data at all. The CTCF_disc1.POUR seems also to score high with Energy_AUC, so it’s not clear that the MNCP is the only factor of relevance here. The CTCF.1_5.ZLAB seems to be most affected by the Energy vs GOMER scoring, and not the MNCP approach. Even if these issues were resolved, it is impossible to know whether these motifs are ‘generally long or with high IC’ from Table 2, because the other motifs aren’t shown. It would be much clearer if the mean and variance of the length & IC for all motifs were also provided for context, or even better correlate the relative score AUC to MNCP differences by length or IC, to truly see if a trend exists.

Figure 6.
Please clarify in the figure legend whether these values are for averages over multiple ChIP datasets (as was discussed above), and if so how these averages are determined.

“Maximum and sum log-odds scoring had low discriminative power for most of the motifs when all three statistical measures are used (Figure 6)”. What are the three statistical measures you’re referring to, and where’s the data? I only see data for AUC. Please clarify.

Table 3. Please be explicit in the figure legend about what the ‘Mean’ and ‘Median’ refer to (i.e., mean and median AUC values calculated over X single motif analyses described in Figure 6)

Figure 7/ page 10. “The variation in the scores is particularly reduced when MNCP statistic is used (Figure 7)”. How am I supposed to see this? What is a significant difference in MNCP and how does it compare to a difference in AUC. Based on the coloring scheme presented the results in Figure 6 and Figure 7 look very similar- it is not clear at all that there is any qualitative difference between these two figures except for the different measures used (i.e., an appropriate normalization might make them near equivalent).

Figure 8. It is not clear (nor mentioned) what is being shown in this figure. I assume – but I could be wrong – that we’re looking at AUC values for each factor (i.e., Mef2a etc) averaged over some ChIP-seq datasets, but how are these being compared to each other? Further, how is Motif_IC which is a function just of the PWM being compared to a scoring function. I can’t speak to the conclusions being reached as I don’t currently know what data is being shown. Much more clarification is warranted in the text and figure

caption.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Author Response 10 Feb 2016

Caleb Kipkurui, Rhodes University, South Africa

Thank you very much for taking the time to review our paper and provide recommendations.Your comments have been very helpful in improving the paper.

Table 1
Corrected

Methods/Data
On repeat masked sequences, we have updated the paper to clarify the we did not include any sequences in test or negative sequences that contained masked positions. By a similar set of negative sequences, we mean matched in length and number of sequences.

Figure 3/Results
The figure caption is updated to clarify. The number of the TFs used was decided based on the availability of ChIP-seq data as well as having motifs in more than 10 of the databases used. A list of the motifs used is provided in the data repository as well as the specific ChIP-seq data used. See also Methods / Data paragraph 3.

Page 6:
On cell line specific binding, an average of the scores of all the available cell lines should be used in motif assessment. We have updated the statement for clarity.

Figure 4
We have changed to using results from GOMER scoring since they are similar; the effect described is only pronounced in Energy scoring.

Page 6/Figure 4

We had incorrectly mentioned the wrong motif to be significantly enriched. We have corrected this and also provided further evidence to the effect that the cell line used in the assessment does actually have an effect on the ranking of the motifs. The conclusions remain valid.

Page 8/Figure 5 (now Figure 6).

We acknowledge that the figure we had used did not present the intended information correctly. We changed the figure to present the general information on the effect of statistics on the ranking of the motifs. We observe that, when normalized, the MNCP and AUC scores do not differ, except for slight difference in some TFs like Hnf4a, Ctcf, Gata3. However, the Pearson and Spearman's correlation scores vary greatly. The plot of the standard deviation of scores as

represented by error bars in Figure 6 demonstrates why we consider correlation scores to be reliable than the other scores. We have added clarification of this point to the paper. Thank you again for pointing out the problem.

Figure 6 (now Figure 7)

The caption has been updated for clarity.

Table 3 (Now Table 2)

Clarified

Figure 7 (now Figure 8)/page 10

The figure has been updated to include information on correlation statistics.

Figure 8 (now Figure 9)

Our apologies for the lack of detail. The figure caption has been updated for clarity. We correlate the scores for the various motifs (for each scoring function) to the length and information content of the motifs to determine whether the scores obtained are in any way influenced by the motif characteristics.

Competing Interests: No competing interests were disclosed.