PLOS | ONE

RESEARCH ARTICLE

Identification of Predictive Cis-Regulatory Elements Using a Discriminative Objective Function and a Dynamic Search Space

Rahul Karnik[1¤], **Michael A. Beer**[1,2]*

[1] Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, United States of America, [2] McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD, United States of America

¤ Current Address: Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, United States of America
* mbeer@jhu.edu

Citation: Karnik R, Beer MA (2015) Identification of Predictive Cis-Regulatory Elements Using a Discriminative Objective Function and a Dynamic Search Space. PLoS ONE 10(10): e0140557. doi:10.1371/journal.pone.0140557

Editor: Martina Stromvik, McGill University, CANADA

Received: February 9, 2015

Accepted: September 28, 2015

Published: October 14, 2015

Data Availability Statement: Source code for MotifSpec and documentation describing installation and operating instructions are available from our website: http://www.beerlab.org/motifspec and at https://github.com/rakarnik/motifspec.

Funding: M. Beer was supported by the Searle Scholars program and NIH grant HG007348. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

The generation of genomic binding or accessibility data from massively parallel sequencing technologies such as ChIP-seq and DNase-seq continues to accelerate. Yet state-of-the-art computational approaches for the identification of DNA binding motifs often yield motifs of weak predictive power. Here we present a novel computational algorithm called MotifSpec, designed to find predictive motifs, in contrast to over-represented sequence elements. The key distinguishing feature of this algorithm is that it uses a dynamic search space and a learned threshold to find discriminative motifs in combination with the modeling of motifs using a full PWM (position weight matrix) rather than *k*-mer words or regular expressions. We demonstrate that our approach finds motifs corresponding to known binding specificities in several mammalian ChIP-seq datasets, and that our PWMs classify the ChIP-seq signals with accuracy comparable to, or marginally better than motifs from the best existing algorithms. In other datasets, our algorithm identifies novel motifs where other methods fail. Finally, we apply this algorithm to detect motifs from expression datasets in *C. elegans* using a dynamic expression similarity metric rather than fixed expression clusters, and find novel predictive motifs.

Introduction

Multiple mechanisms exist to modulate protein levels in a cell and create a dynamic cellular phenotype from a static genotype. One such mechanism is transcriptional regulation. Transcription factors (TFs) bind to intergenic cis-regulatory elements and enhance or inhibit the transcription of their target genes. Identifying the DNA binding specificities of transcription factors is necessary to decipher the regulatory network in the cell, identify disease causing mutations in these elements, and engineer synthetic organisms to perform specific biochemical functions. Several technological platforms can be used to identify the binding specificities of the DNA-binding domains of transcription factors to cis-regulatory elements in DNA. Binding can be directly measured in vivo using chromatin immunoprecipitation followed by microarray hybridization (ChIP-chip) [1] or sequencing (ChIP-seq) [2], or in vitro using protein binding microarrays (PBMs) [3]. Alternatively, gene co-expression implies the binding of TFs to cis-regulatory elements, and thus indirectly indicates binding by a common regulator. The spatial resolution of these technologies are typically unable to resolve short TF binding sites, so motif-finding algorithms are usually used to identify TF binding specificity within the set of longer experimentally determined regions [2–4].

Despite these significant advances in technology, there is still a substantial gap in our ability to generate PWMs that accurately describe binding specificities from these experiments. For instance, 246 candidate DNA-binding proteins from yeast were assayed with PBMs [5]. Of these, predictive motifs were found in only 89 cases, or 36% of the factors assayed. Similarly, 23 transcription factors from *Caenorhabiditis elegans* were assayed using ChIP-seq as part of the modENCODE project [6]. Predictive motifs were found for only 8 (35%) of these factors. Considering the rapid growth of these technologies [6–9], improved algorithms to extract regulatory sequence information from these data sets would be of clear value. In particular, better motif models could improve the identification of regulatory mutations associated with common human disease [10] and could be used to develop improved techniques to detect regulatory variation responsible for differences in gene expression between species [11–13].

Generally, motif-finding algorithms search a set of sequences for shared cis-regulatory elements. We term this set of sequences the search space. Early motif-finding algorithms optimized for over-represented sequence motifs, which are sequence patterns found more often in the search space than would be predicted by a null or background sequence model. Successful algorithms of this class include AlignACE [4] and MEME [14], which use Gibbs sampling and expectation maximization respectively, to search for the optimal sequence motif. A discriminative approach, in contrast, searches for specific motifs, those cis-regulatory elements which are present at a higher frequency in a positive set of sequences than in a negative set of sequences. A frequency difference in the positive and negative sets is required for accurate classification of one set from the other, and is often not reflected by over-representation in the positive set, as many instances of the over-represented motifs are also found in the negative set. Several algorithms of the discriminative type have been developed previously, including Amadeus [15], DREME [16], HOMER [17], Dimont [18], DECOD [19], and others [20–23]. By necessity, a discriminative objective function is more expensive to compute; it requires scoring not only the sequences in the positive search space, but also those sequences in the negative set to establish the scarcity of the motif in negative sequences. Consequently, some previous algorithms have used a *k*-mer sequence model for the motif while performing discriminative motif finding. Our algorithm, which we call MotifSpec, uses a full position-weight matrix as its sequence model, which we show performs better than many existing discriminative models, and is comparable in accuracy to HOMER [17] and Dimont [18].

The positive set can be defined through one of three methods. First, we might have direct binding data for the protein or TF, and the set of sequences whose binding score is above a threshold is used as the positive set. This threshold can be a simple rank, or can result from computation of a p-value given some null binding model. This approach is common with ChIP-chip, ChIP-seq and PBM experiments. Second, we might use co-expression as a proxy for binding, and choose a correlation threshold to define the search space. This method often involves clustering co-expressed genes and then running a motif-finding algorithm on the upstream regions of genes in the individual cluster [4,24–27]. Finally, we might use prior biological annotation to identify bound sequences. In the first two cases, the optimal boundary between the positive and negative sets is generally not obvious, and changing the boundary threshold will change the membership of the positive and negative sets.

Most existing discriminative motif finding algorithms consider the positive and negative search spaces to be fixed. Instead of using a fixed set of positive sequences, a dynamic approach allows the boundary between the positive and negative sequences to evolve during the search procedure. A dynamic threshold can be applied when searching for motifs in any set of continuous enrichment data, such as ChIP-seq peak intensity. MotifSpec uses such a dynamic search space to optimize the specificity of the motif, which we show improves quantitative measures of motif predictive power on PBM and expression datasets.

Methods

MotifSpec is a heavily modified version of the Gibbs sampling algorithm implemented in AlignACE [26]. The key innovations in MotifSpec are a dynamic search space, a dynamic threshold for sequence score, and a hypergeometric discriminative objective function.

Dynamic threshold to determine positive and negative sets

As mentioned above, current algorithms typically use a fixed search space. In contrast, MotifSpec dynamically optimizes the threshold defining the positive and negative sets using the objective function discussed below. In the case of TF binding, this threshold is a binding intensity. This dynamic search space is appropriate given that TFs have widely varying target set sizes.

Dynamic threshold for sequence score

In both AlignACE and MotifSpec, sequence positions are scored according to the equation: $L = \Pr(S|\theta)/\Pr(S|\theta_0)$, where L is the site score or likelihood ratio, S is the site being scored, θ is a PWM sequence model, and θ_0 is a background distribution. This odds-ratio is then converted into a probability using a Bayesian framework. We compute the probability of the PWM model θ given the site S currently being scored as: $\Pr(\theta|S) = L \Pr(\theta)/(L \Pr(\theta) + 1 - \Pr(\theta))$.

In AlignACE, a fixed $\Pr(\theta|S)$ threshold determines whether a specific site is an instance of the current PWM, and whether it should be added to the PWM for the next iteration. Instead of a fixed threshold, MotifSpec dynamically adjusts this threshold to maximize the objective function discussed below. We call this threshold the "sequence threshold" because it determines how high the sequence of a site must score by the current PWM to be considered an actual binding site, and to be added to the model. In addition, AlignACE has a parameter called *expect*, which is the prior for number of instances of the motif in the search space. This parameter is used to calculate $\Pr(\theta) = (ew + x(1 - w)) / T$, where e is expect, x is the current number of motif instances, w is a weight assigned to the prior and T is the total number of positions available in the search space. Instead of relying on the difficult to estimate expect parameter, MotifSpec assumes one motif instance per sequence, replacing e with the number of sequences in the current search space, s_1.

Model components and objective function

In order to support the dynamic thresholds for sequence score and search space membership, the MotifSpec motif model consists of a PWM and two additional components:

1. Search space threshold: the minimum binding score that a sequence must have to be included in the search space or positive set. The set of sequences that are above this threshold is the set s_1. In the case of ChIP-seq, this threshold score is read depth; for PBM, it is the binding intensity of the oligo; in the case of expression similarity, it is a correlation measure.

2. Sequence score threshold: the minimum site score that a site in a sequence must achieve to be considered an instance of the motif. The set of sequences that have a site scoring above this threshold is the set s_2.

MotifSpec uses a discriminative objective function called a specificity score, which measures the enrichment for sequences to be in both s_1 and s_2. Given x sequences that are in the intersection of the above sets, and N total sequences in the positive and negative sets, the specificity score is defined using the hypergeometric distribution as

$$\text{Specificity score} = -\log\left(\sum_{i=x}^{\min(s_1,s_2)} \frac{C(s_1, i)C(N - s_1, s_2 - i)}{C(N, s_2)}\right)$$

This specificity score is the negative logarithm of the group specificity score used by Hughes et al [4]. Since N is typically large for genomic datasets, we calculate the summation in the specificity score in log space using $\log(x + y) = \log(x) + \log(1 + \exp(\log(y) - \log(x)))$, where x and y are individual probability terms for different values of i in the equation above.

Weighted PWM

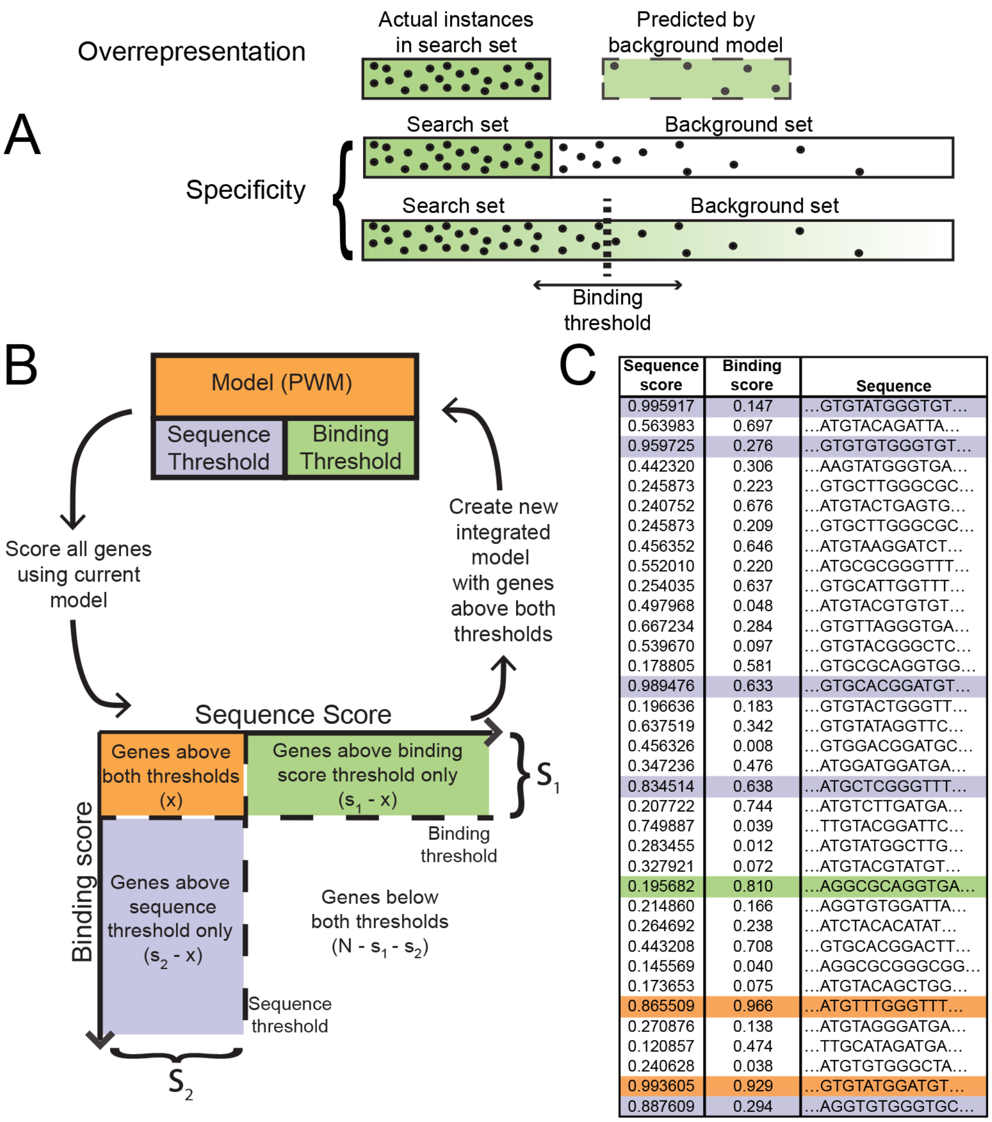
AlignACE and most other motif finders use an equal contribution from each motif instance to compute the PWM. MotifSpec weights the contribution of each instance to the PWM according to the binding score of that instance. Say that we have n instances of a motif. Let $\mathbf{w}$ be the vector of binding scores of the sequences normalized to be between 0 (low) and 1 (high). Assume that the PWM has k columns and that Ii,j,b is the indicator variable of having the base b at position j in instance i of the motif. Then the probability fj,b of having base b at position j in the motif in an unweighted PWM is $f_{j,b} = (\sum_{i=1}^{n} I_{i,j,b})/n$. Instead, MotifSpec calculates the weighted PWM using $f_{j,b} = (\sum_{i=1}^{n} I_{i,j,b}w_i)/(\sum_{i=1}^{n} w_i)$.

Algorithm

The MotifSpec algorithm iteratively optimizes the PWM and the thresholds in the model (Fig 1). MotifSpec initializes the model by choosing a random site from the positive search space. Similar to AlignACE, convergence is measured by improvement in the specificity score; MotifSpec stops iterating after a series of iterations without improvement have occurred, with the number of iterations being a customizable parameter. MotifSpec then alternately adjusts the binding score threshold and sequence score threshold to maximize the objective function, given the current PWM. Sites are rescored using these new thresholds. With the new thresholds, the iteration process is repeated, again until *minpass* iterations without improvement occur, and the current model is output. After the first motif is found, subsequent searches are performed with new random starts, terminating early if the current model is similar (CompareACE score greater than a threshold, default 0.9) to a motif previously found with a higher specificity score. The number of such restarts is $s_1/(w\ k)$, where $s1$ is the size of the search space (in base pairs), w is the number of columns in the PWM motif model, and k is a sampling parameter.

Other Algorithms

All algorithms were run in discriminative mode and with default parameters unless otherwise noted here. HOMER version 4.7 was downloaded from http://homer.salk.edu/homer/download.html and run with motif lengths 8,10,12,14,16,18, and 20, and the highest scoring motif is used for comparison. Dimont was downloaded from http://www.jstacs.de/index.php/


A
Overrepresentation
Actual instances in search set
Predicted by background model
Specificity
Search set
Background set
Search set
Background set
Binding threshold
B
Model (PWM)
Sequence Threshold
Binding Threshold
Score all genes using current model
Create new integrated model with genes above both thresholds
Sequence Score
Binding score
Genes above both thresholds (x)
Genes above binding score threshold only (s_1 - x)
s_1
Binding threshold
Genes above sequence threshold only (s_2 - x)
Genes below both thresholds (N - s_1 - s_2)
Sequence threshold
s_2
C

Sequence score	Binding score	Sequence
0.995917	0.147	…GTGTATGGGTGT…
0.563983	0.697	…ATGTACAGATTA…
0.959725	0.276	…GTGTGTGGGTGT…
0.442320	0.306	…AAGTATGGGTGA…
0.245873	0.223	…GTGCTTGGGCGC…
0.240752	0.676	…ATGTACTGAGTG…
0.245873	0.209	…GTGCTTGGGCGC…
0.456352	0.646	…ATGTAAGGATCT…
0.552010	0.220	…ATGCGCGGGTTT…
0.254035	0.637	…GTGCATTGGTTT…
0.497968	0.048	…ATGTACGTGTGT…
0.667234	0.284	…GTGTTAGGGTGA…
0.539670	0.097	…GTGTACGGGCTC…
0.178805	0.581	…GTGCGCAGGTGG…
0.989476	0.633	…GTGCACGGATGT…
0.196636	0.183	…GTGTACTGGGTT…
0.637519	0.342	…GTGTATAGGTTC…
0.456326	0.008	…GTGGACGGATGC…
0.347236	0.476	…ATGGATGGATGA…
0.834514	0.638	…ATGCTCGGGTTT…
0.207722	0.744	…ATGTCTTGATGA…
0.749887	0.039	…TTGTACGGATTC…
0.283455	0.012	…ATGTATGGCTTG…
0.327921	0.072	…ATGTACGTATGT…
0.195682	0.810	…AGGCGCAGGTGA…
0.214860	0.166	…AGGTGTGGATTA…
0.264692	0.238	…ATCTACACATAT…
0.443208	0.708	…GTGCACGGACTT…
0.145569	0.040	…AGGCGCGGGCGG…
0.173653	0.075	…ATGTACAGCTGG…
0.865509	0.966	…ATGTTTGGGTTT…
0.270876	0.138	…ATGTAGGGATGA…
0.120857	0.474	…TTGCATAGATGA…
0.240628	0.038	…ATGTGTGGGCTA…
0.993605	0.929	…GTGTATGGATGT…
0.887609	0.294	…AGGTGTGGGTGC…

Fig 1. MotifSpec optimizes for specificity rather than over-representation and uses a dynamic search space. (A) An over-represented motif is found in the search space more often than expected according to some background model. It is not necessarily predictive. A specific motif is found in a much higher frequency in the search space than in the background sequences. A dynamic search space threshold finds the optimal search space such that the motif is most discriminative. (B) A schematic of the MotifSpec algorithm. The PWM model is initialized with a random sequence and position in the search space. The model is iteratively refined and the motif and binding score thresholds are adjusted at convergence to maximize specificity. (C) An example of sequences scored using the model. Each sequence has a motif score and a binding score. The binding score determines if a sequence is in the search space. The motif score determines if the sequence has an instance of the motif. The sequences are color-coded according to the set to which they belong as defined in (B).

doi:10.1371/journal.pone.0140557.g001

Dimont/Download, and was run in discriminative mode by giving each positive sequence a "signal" of 1000, each negative sequence a "signal" of 0, and used a "peak" in the center of each sequence. DECOD version 1.01 was downloaded from http://sb.cs.cmu.edu/DECOD/, and was run with motif width 10 and 12, as larger width required excessive memory (>100GB). Amadeus was downloaded from http://acgt.cs.tau.ac.il/amadeus/. DREME and MEME were run from the MEME suite version 4.10.1 downloaded from http://meme-suite.org/. Weeder was downloaded from the ModTools site (http://159.149.160.51/modtools/). Seed-and-Wobble motifs were downloaded from Uniprobe [9]. Only MEME, Weeder, and AlignACE were run in non-discriminative mode.

Human, Mouse, and Worm ChIP-seq datasets

We analyzed 16 mammalian ChIP-seq datasets. Three of the datasets (CTCF [27], NRSF [2], and the estrogen receptor (ER) [28]) measured binding in human cells, while the other 13 measured binding of mouse TFs in embryonic stem cells [29]. For the human TFs, we downloaded the raw data from the Gene Expression Omnibus (GEO) database. We re-processed the raw data using the MACS algorithm [30] with default parameters and designated high confidence peaks (false discovery rate $< 0.01\%$) as the positive set. This process resulted in positive sets with 5444, 2417 and 1225 peaks for CTCF, NRSF and ER respectively. A 300 bp window of genomic sequence around each peak was used for analysis for ER and CTCF, a 500bp was used for NRSF. For mouse embryonic stem cell datasets, we used the set of bound sequences defined in [29] as the positive set. We also analyzed ChIP-seq data for *C. elegans* from the modENCODE project [6]. We used the peaks designated “appropriate for downstream analysis" in [31] as positive sets. For each positive set, we created a negative set using a previously published procedure that matched the sequence lengths, GC content and repeat fraction of the positive set [32]. We have the freedom to select larger negative sets which usually lead to more predictive motifs. For the human and mouse datasets we used 2x each positive set and for worm used 4x each positive set.

Yeast PBM datasets

We downloaded data from 132 PBM experiments with 89 yeast TFs from Uniprobe [9]. The raw scores for the approximately 40,000 60-mer probes were translated to strictly positive values and fitted with a log-normal null distribution. The goodness of fit of the scores to the model was verified with qq-plots (S1 Fig). We used 1-F as a normalized (0; 1) binding score for each probe, where F is the cumulative distribution function of the log-normal translated scores. When learning motifs with MotifSpec, we dynamically optimized the positive and negative sets. For subsequent ROC curve analysis to compare the predictive power of MotifSpec motifs to previous approaches, we used normalized binding scores greater than 0.9 to define the positive set, and binding scores less than 0.5 for the negative set. Our comparative analysis is insensitive to these cutoffs.

Synthetic sequence-expression data

To test our dynamic expression clustering, we constructed a synthetic dataset consisting of 5000 sequences with lengths sampled from a Gaussian distribution with mean 800bp and standard deviation 100bp, and which have the same GC content as yeast intergenic sequences. Instances of test motifs were seeded into these sequences. To approximately reproduce the structure of yeast intergenic regions, we seeded two classes of motifs: "functional" and "non-functional". Four "non-functional" motifs were seeded into each sequence at random, but the "functional" motifs were only seeded into co-expressed sets of genes. These sets were of varying size, and one "functional" motif was seeded into each member of a co-expressed gene set. A total of 80 functional motifs were seeded. To mimic the dominant yeast expression patterns of stress-induction and stress-repression, 40 of the functional motifs were seeded into stress-induced gene sets, and 40 were seeded into stress-repressed gene sets. Individual expression patterns were sampled from a Gaussian distribution with means centered on these two anticorrelated expression patterns. This procedure produced pairwise correlations between expression pattern means that matched the actual expression data described below.

Yeast and *C*. *elegans* sequence-expression datasets

We extracted upstream sequences for all yeast (*Saccharomyces cereviseae*) ORFs as previously described [25]. We also created a combined gene expression dataset from three different yeast studies, which included cell cycle timepoints and various metabolic stimuli, for a total of 5228 genes across 292 conditions [33–35]. Similarly, we extracted upstream sequences for all genes in *C*. *elegans* as previously described [25] and combined expression data from three different studies in *C*. *elegans* [36–38] for a total of 82 conditions and 5691 genes.

Evaluation of motifs for ChIP-seq and PBM datasets

To measure the predictive power of our motifs and compare to those previously reported, we scored the positive and negative sets with each PWM using ScanACE, and ranked all sequences according to the highest scoring site in each sequence. Using this ranked list, we plotted a receiver operator characteristic (ROC) curve and used the area under the ROC curve (auROC) as a measure of how well a motif was able to discriminate between the positive and negative sets.

Background model

AlignACE uses a single nucleotide frequency model (equivalent to a 0th order Markov model) to calculate the site score. Later algorithms (e.g. [14]) have shown that the use of a higher order background model can prove beneficial. In keeping with this trend, MotifSpec can use a background Markov model with order up to 5. In practice, we tend to use a 3rd order Markov model. Increasing the order of the background model did not result in consistent improvement in performance (S2 Fig). Since our objective function penalizes motifs according to their actual frequency in the negative/background set, it is likely that the background model, which is a summary statistic of the background set, is not as important to performance.

Gapped motifs and number of PWM columns

One advantage that AlignACE provides over competing algorithms is the ability to find gapped motifs. The PWM model is not necessarily contiguous bases and can actually include gaps. For example, the width of the motif might be 15 bases, but only 10 of these bases might be informative and therefore included in the PWM. This gapped PWM model is useful for TFs that bind as a dimer to two sets of constrained DNA bases separated by unconstrained bases. For example, the yeast TF Gal4 binds to the sequence pattern CGGnnnnnnnnnnnGGC. MotifSpec takes the gapped motif concept one step further: it has the ability to add or remove columns to the PWM randomly during the motif search to see if it improves the specificity score. We use this setting for genome-wide searches using expression data, since we are looking for novel motifs that may have an unpredictable number of informative columns.

Performance heuristics

Given the definition of the specificity score, it is obvious that each iteration of the MotifSpec algorithm requires the scanning of *N* total sequences for motif instances, unlike over-representation based algorithms such as AlignACE and MEME which would scan the search space alone. Since each random start of the MotifSpec algorithm is independent, however, we can parallelize the searches and essentially reduce running time by a factor equaling the number of parallel instances of MotifSpec (S3 Fig).

Another heuristic used to increase scanning performance is the maintenance of a list of the highest scoring site in each of the input sequences. Since only the highest scoring site determines membership in the set of sequences containing an instance of the motif, we do not need to scan every position of every sequence. Instead, we only scan the position that was the highest scoring site in each sequence in the previous iteration. If the PWM has changed considerably from the previous iteration, MotifSpec then triggers a full scan of all positions in all sequences.

Software availability

Source code for MotifSpec and documentation describing installation and operating instructions are available from our website: http://www.beerlab.org/motifspec and at https://github.com/rakarnik/motifspec.

Results

Human ChIP-seq datasets

We first used ChIP-seq data for three human transcription factors: CTCF, NRSF, and the estrogen receptor (ER), to compare the performance of MotifSpec to five other algorithms: DREME [16], Amadeus [15], HOMER[17], Dimont[18], and DECOD[19]. Positive and negative sets were identical for each method and chosen as described in Methods. In all three cases, the top motif found by MotifSpec was able to discriminate significantly better between the positive and negative sets (higher auROC) than DREME and Amadeus and DECODE, and was generally similar or marginally better than HOMER and Dimont (Fig 2, S4–S6 Figs) The motifs found by each algorithm are shown in Fig 2E. MotifSpec consistently found longer motifs than either DREME or Amadeus. As these algorithms are word-based, we speculate that finding high-scoring exact matches to individual long *k*-mers or regular expressions is less likely and that these long words get filtered out at an early stage in the algorithm. In contrast, as an entirely PWM-based algorithm, MotifSpec is able to detect these longer motifs with gaps and degenerate positions. PWM models typically have more parameters than word based models. To ensure that the better discriminative power of the motifs found by MotifSpec was not due to over-fitting, we performed 5-fold cross-validation on the three datasets using MotifSpec. The motifs found were almost identical to those found using each complete dataset, and the auROC for the test sets were consistently within 1% of the auROC found on the whole dataset.

Mouse ChIP-seq datasets

1We next ran MotifSpec on 13 TF ChIP-seq data sets generated in mouse embryonic stem cells [29], and compared our motifs to the previously published motifs found using DREME [16]. Motifs found by MotifSpec were similar, but not identical, to those found by DREME in each dataset. We evaluated the ability of the top motif in each dataset to recover positive sequences from each dataset as above using auROC. In 10 of the 13 datasets, the top motif found by MotifSpec had higher auROC than the top motif found by DREME (Fig 3). In the three remaining cases (Zfx, Klf4, Esrrb), the auROCs were almost identical. For some cases, the auROC values for the best single motifs are low, indicating more combinatorial regulation by those factors.

C. *elegans* ChIP-seq datasets

We ran MotifSpec on the ChIP-seq data for 23 worm transcription factors [31]. We found four additional motifs, for ELT-3, GEI-11, LIN-15B, and PQM-1, which were not reported in the

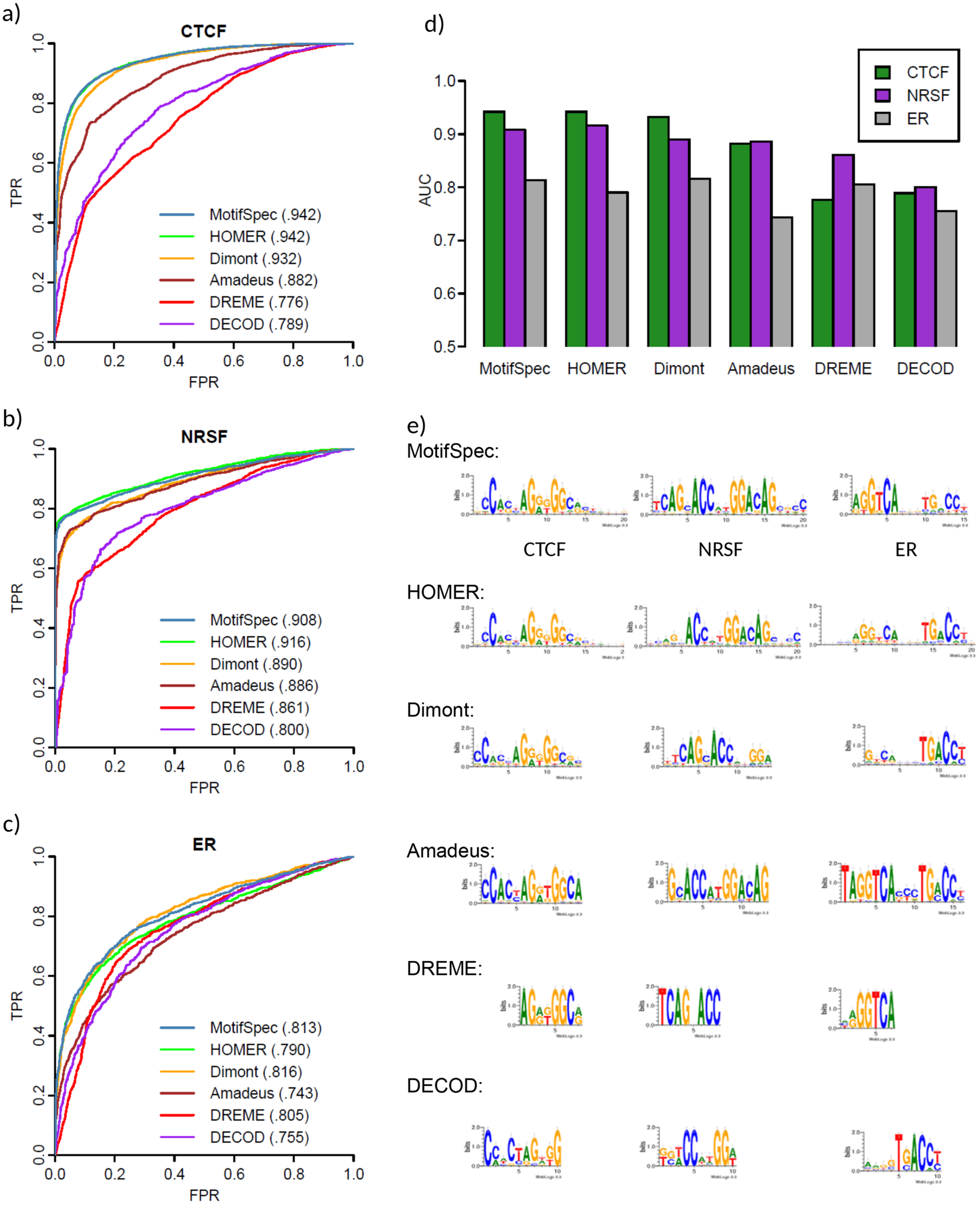

a)
CTCF
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0
MotifSpec (.942)
HOMER (.942)
Dimont (.932)
Amadeus (.882)
DREME (.776)
DECOD (.789)
b)
NRSF
MotifSpec (.908)
HOMER (.916)
Dimont (.890)
Amadeus (.886)
DREME (.861)
DECOD (.800)
c)
ER
MotifSpec (.813)
HOMER (.790)
Dimont (.816)
Amadeus (.743)
DREME (.805)
DECOD (.755)
d)
AUC
0.5
0.6
0.7
0.8
0.9
1.0
🟩 CTCF
🟪 NRSF
▨ ER
MotifSpec
HOMER
Dimont
Amadeus
DREME
DECOD
e)
MotifSpec:
CTCF
NRSF
ER
HOMER:
Dimont:
Amadeus:
DREME:
DECOD:

Fig 2. Human ChIP-seq results. MotifSpec performs comparably to HOMER and Dimont and consistently better than DECOD, DREME, and Amadeus in finding a discriminative motif when run on ChIP-seq data for three human transcription factors, CTCF, NRSF and the estrogen receptor (ER). Panels a, b, and c show the ROC curves and auROC values for the top scoring motif from each program when run on the three datasets. Panel d shows a summary comparison of auROC for each algorithm and motif, and panel e shows the top scoring motif found by each program.

doi:10.1371/journal.pone.0140557.g002

original analysis (Fig 4). The motif found for LIN-15B was also one of the top motifs found by dynamic expression clustering as described below.

Yeast PBM datasets

We next used MotifSpec to find motifs in the yeast PBM data [5]. We ran MotifSpec using the dynamic search space mode, using normalized binding scores as described in Methods. Using the list of probes ranked by sequence score, we plotted the ROC curve for each TF-microarray pair and calculated auROC. Using this benchmark, we compared the motifs found by MotifSpec to the motifs reported in the original study as found by the Seed and Wobble algorithm [3,5]. We excluded 41 of the 132 experiments where either auROC was less than 0.75, thereby eliminating any experiments where neither algorithm found a sufficiently predictive motif. As shown in Fig 5, the motifs found by MotifSpec outperformed those found by Seed-and-Wobble in 76 of the 91 experiments (83%). This performance improvement was consistent, regardless of the p-value threshold used to define the positive set. With p-value thresholds of 0.05 and 0.01, motifs found by MotifSpec were more predictive in 83/104 (80%) and 93/119 (79%) experiments respectively (S7 Fig).

Dynamic expression clustering

Previous studies have attempted to detect regulatory elements by identifying sets of co-expressed genes and searching for shared sequence motifs in the upstream regions of these sets of genes [4,25,26]. These efforts first clustered genes by their expression, then used motif-

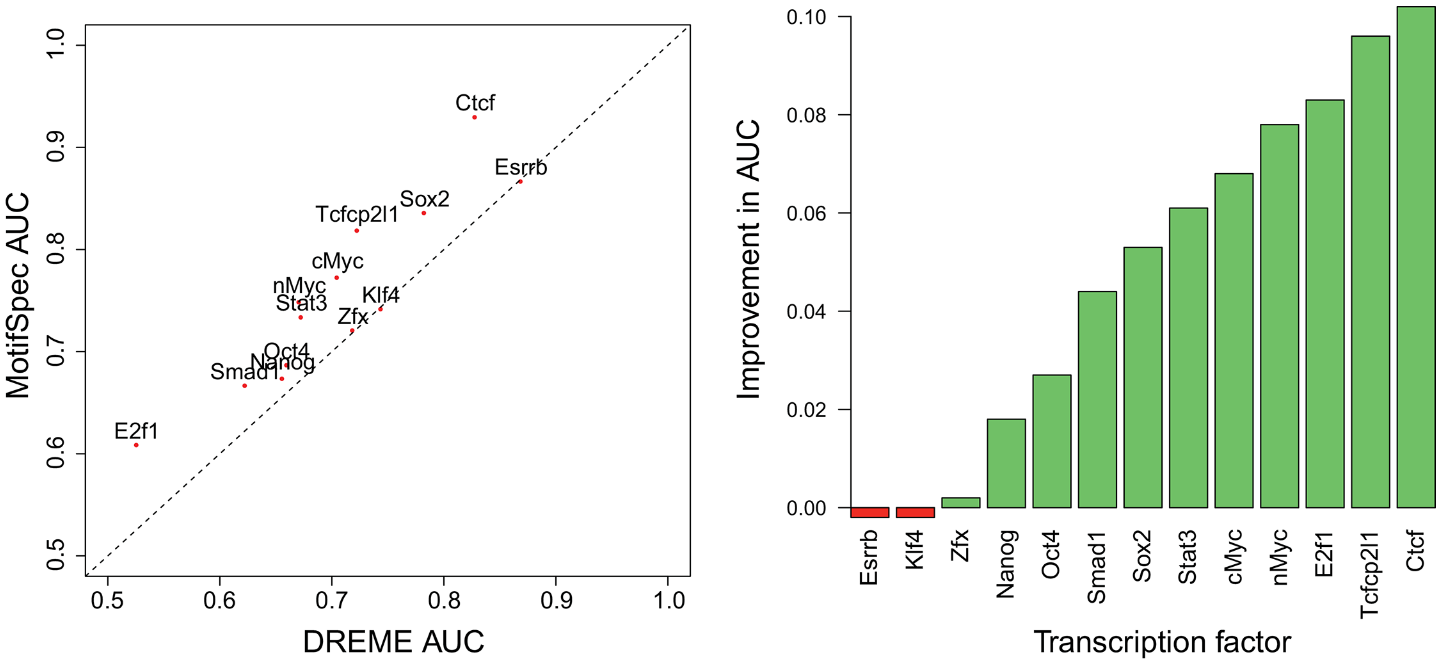


Fig 3. Mouse ChIP-seq results MotifSpec outperforms DREME when run on ChIP-seq data for 13 transcription factors from mouse embryonic stem cells. The left panel shows a plot of the AUC for the top motif reported by MotifSpec against the AUC for the top motif reported by DREME, while the right panel shows the improvement in AUC for the MotifSpec motif relative to the DREME motif.

doi:10.1371/journal.pone.0140557.g003

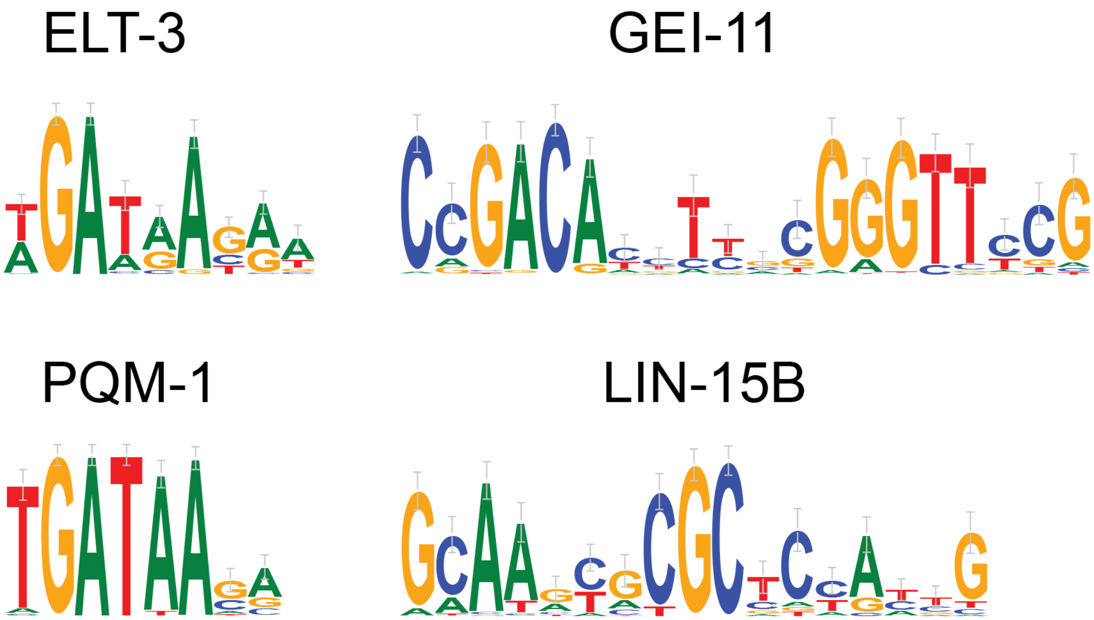


Fig 4. modENCODE ChIP-seq results. Binding specificities for four *C*. *elegans* transcription factors as learnt from ChIP-seq data from the modENCODE project.

doi:10.1371/journal.pone.0140557.g004

finding algorithms on the upstream regions of each cluster. Given inherent biological and experimental noise, these methods are limited in their ability to tease apart similarly expressed regulatory programs, as similarly expressed genes can be assigned incorrectly to clusters and lower the sensitivity of the subsequent motif-finding step. MotifSpec can increase sensitivity relative to this two-step approach by using its dynamic search spaces to search for elements in co-expressed gene sets without starting with predefined co-regulatory clusters. MotifSpec identifies genes that have similar expression profiles and a shared sequence motif in their upstream regions, and iteratively refines the model of both the expression profile and the sequence motif. This approach was used to search for regulatory elements in genome-wide datasets using combined sequence and expression data. We first test MotifSpec on simulated expression and sequence data, and then on actual yeast expression and sequence data, as it has been emphasized that it is significantly more difficult to detect motifs in actual genomic sequence [39].

Synthetic sequence-expression dataset

To evaluate the performance of this approach on a simulated dataset, we ran MotifSpec against the synthetic data described in Methods. To compare MotifSpec to the two-step approach, we also clustered the genes using k-means clustering on the expression data and ran the motif finding algorithms AlignACE, MEME and Weeder on the upstream sequences of the genes in each cluster. We evaluated the ability of these distinct methods to recover seeded motifs from the upstream sequences. MotifSpec performed much better than the two-step approach, as shown by the precision-recall curve (Fig 6). At a 10% false positive rate (90% precision), MotifSpec recovered 85% of the seeded motifs, while the two-step algorithms recovered between 46% and 60%.

Yeast sequence-expression data

We next compared the performance of MotifSpec to that of a two-step process of clustering using k-means and motif-finding within clusters using AlignACE on the actual yeast

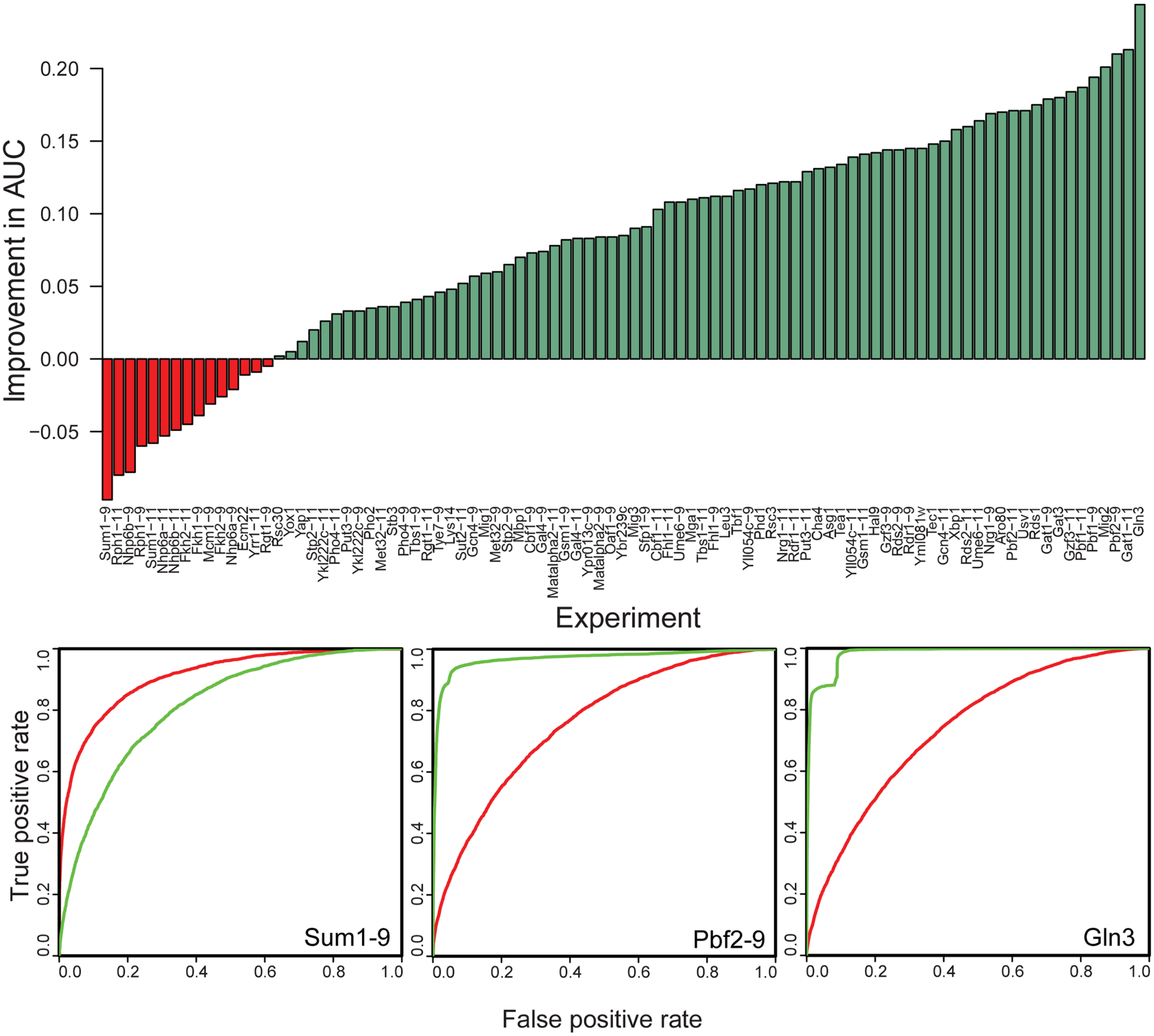


Fig 5. Motifs found by MotifSpec perform better at retrieval of bound probes than the motifs found by Seed-and-Wobble. The barchart shows the percentage improvement in the area under the receiver-operator characteristic (ROC) curve, and the top motif found by MotifSpec performs better than the Seed-and-Wobble motif in the majority of cases where either motif has an AUC of 0.75 or better. Three representative ROC curves are shown, two (Gln3 and Pbf2-9) in which MotifSpec outperforms Seed-and-Wobble and one in which Seed-and Wobble is better (Sum1-9). The red curve is the ROC for the Seed-and-Wobble motif and the green curve is the ROC for the best MotifSpec motif.

doi:10.1371/journal.pone.0140557.g005

expression data described in Methods. The list of motifs generated by each algorithm was compared to a compendium of 97 known yeast motifs. MotifSpec found more known motifs from the expression data (65/97 or 67%) than k-means-AlignACE (39/97 or 40%) at a CompareACE threshold of 0.75 (Fig 7). We also compared the target list for predicted motifs with lists of target genes from yeast ChIP-chip data. More target lists found by MotifSpec overlapped

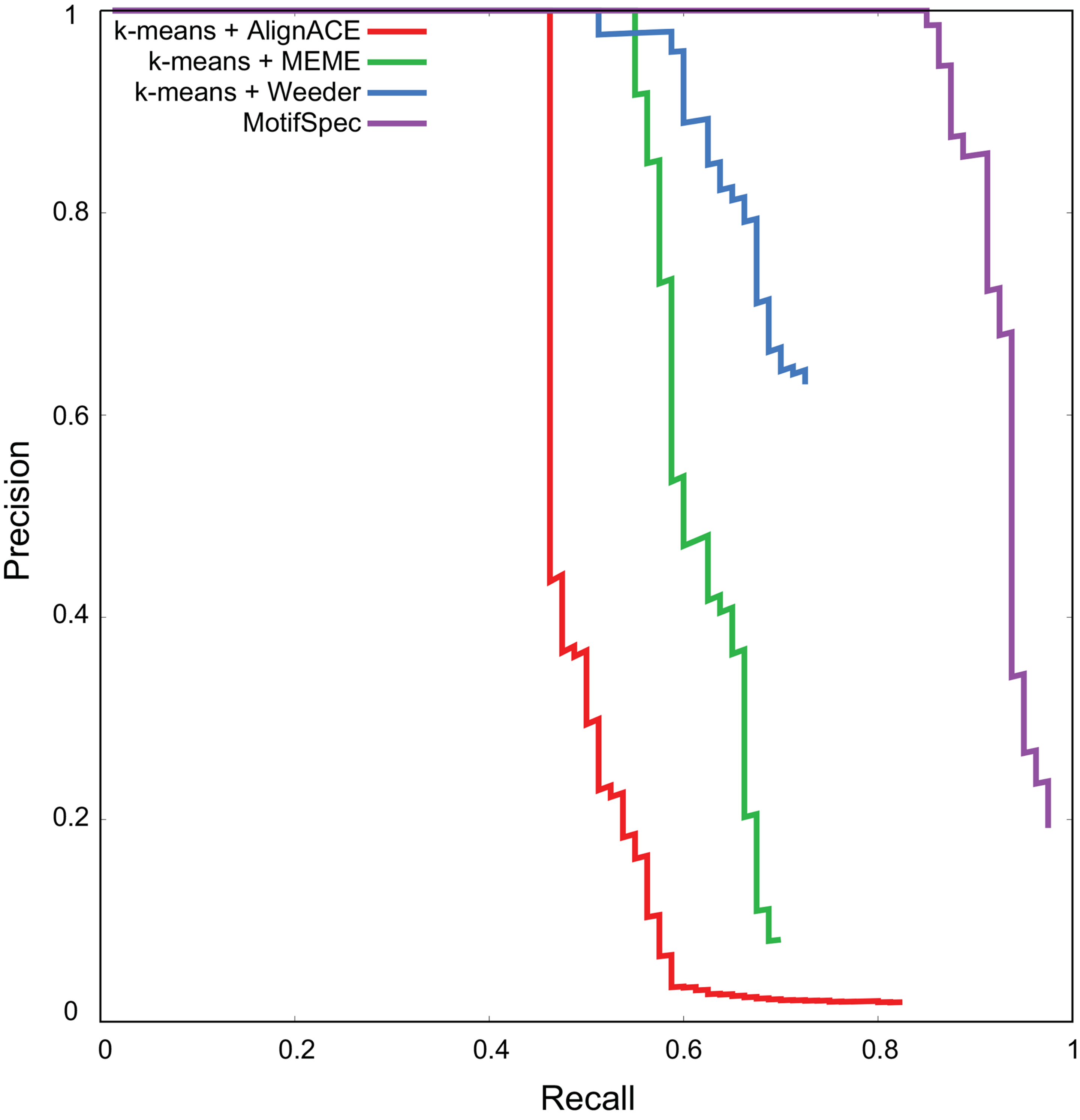


Fig 6. MotifSpec performs better at recovery of seeded motifs from a synthetic sequence-expression dataset than two-step procedures of k-means clustering and motif-finding using AlignACE, MEME and Weeder.

doi:10.1371/journal.pone.0140557.g006

significantly (p-value $< 10^{-7}$) with ChIP-chip target lists than those found by k-means-AlignACE (Fig 7).

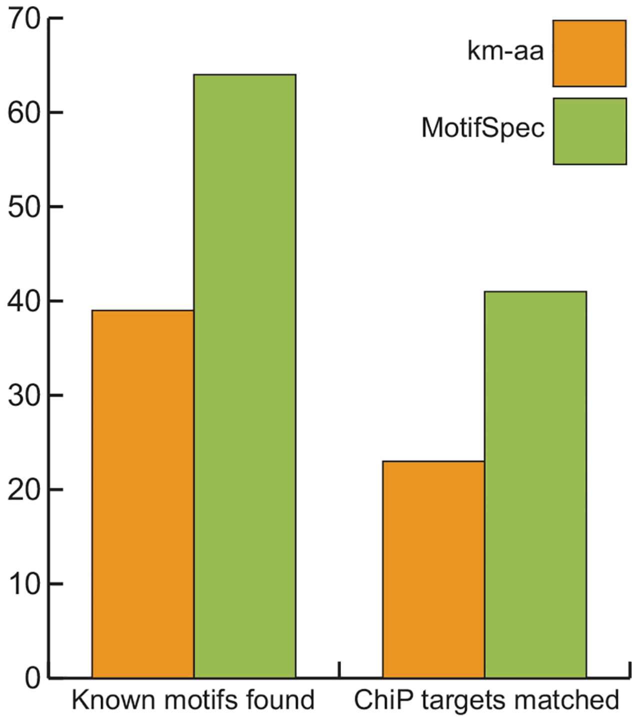


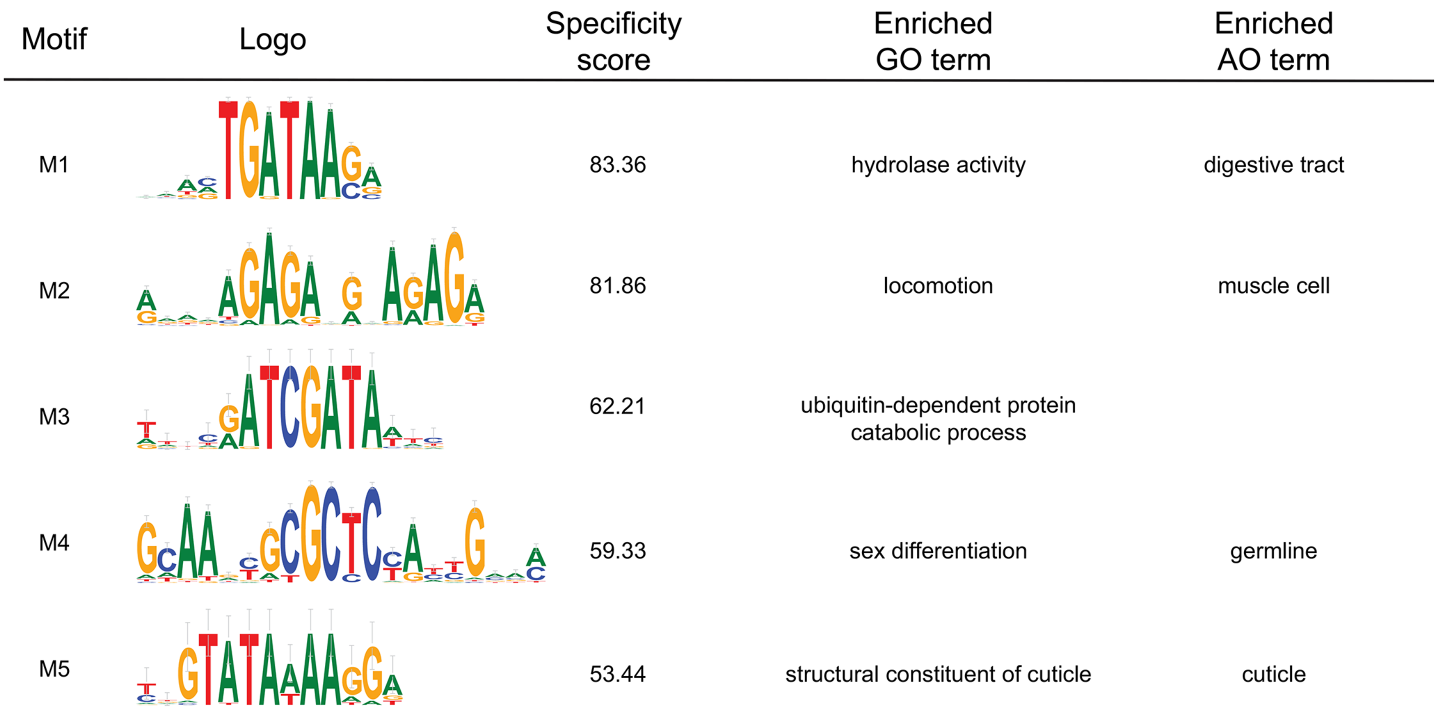
Fig 7. MotifSpec detects more known yeast motifs than the combination of k-means clustering and AlignACE (km-aa). There were 97 known motifs in total. A CompareACE motif similarity score of 0.75 or greater was considered a match. ChIP target sets were considered a match if the hypergeometric p-value for overlap was less than 10^{-7}.

doi:10.1371/journal.pone.0140557.g007

C. *elegans* sequence-expression data

We next used dynamic expression clustering to find motifs using MotifSpec on *C. elegans* expression data. To determine a significance threshold for reporting motifs, we repeated the search on randomized sequences as a negative control. At a specificity score of 26 or higher, we found 135 motifs found in the real dataset and only 10 motifs in the randomized dataset, which translates to a false discovery rate of 7.4%. To generate a non-redundant list, we removed motifs that had a target gene overlap of 30% or greater with a similar motif, leaving 87 motifs total (S1 Table). For each of these motifs, we calculated Gene Ontology (GO) and Anatomy Ontology (AO) [40] enrichment.

The top 5 motifs from this analysis are shown in Fig 8, along with any enriched GO and AO terms. Motif M1 is the known GATA factor binding site. As expected from its intestinal function, GO terms such as "small molecule metabolic process" and "hydrolase activity" are highly enriched in the set of target genes, while the AO terms "digestive tract" and "intestine" are also enriched. Motif M2 was previously found to be associated with the expression of muscle genes [41], and "locomotion" and "muscle cell" are the most enriched ontology terms. We note that several similar GA-rich motifs on our full list (S1 Table) also have high specificity scores and highly overlapping target gene sets, suggesting that the motif may be more degenerate than the highest scoring motif would suggest. Motif M3 is novel and is significantly enriched for "ubiquitin-mediated proteolysis". Its target genes include several F-box family proteins and noncoding RNAs. Motif M4 matches a motif identified as the binding site for CEH-30 from modENCODE ChIP-seq data [6], while our own analysis of the LIN-15B ChIP-seq dataset from the modENCODE project identifies it as well (Fig 4). CEH-30 ensures survival of male-specific



Motif	Logo	Specificity score	Enriched GO term	Enriched AO term
M1		83.36	hydrolase activity	digestive tract
M2		81.86	locomotion	muscle cell
M3		62.21	ubiquitin-dependent protein catabolic process	
M4		59.33	sex differentiation	germline
M5		53.44	structural constituent of cuticle	cuticle

Fig 8. The top 5 motifs found by MotifSpec in a genome-wide search of a *C. elegans* sequence and expression dataset. Alongside each motif is its specificity score and any Gene Ontology (GO) and Anatomy Ontology (AO) terms that were enriched in the list of target genes.

doi:10.1371/journal.pone.0140557.g008

neurons during development [42] and M4 targets are enriched for "sex differentiation", which would support the hypothesis that CEH-30 binds to this motif. LIN-15B is implicated in the development of vulval cells [43]. In either case, it is likely to be functional. Motif M5 is another novel motif and targets "cuticle" genes, including 12 collagen and 3 vitellogenin genes.

Discussion

We have described a novel discriminative motif finding algorithm which uses dynamic search spaces and we evaluated the discovered motifs' predictive performance using ROC analysis. Our algorithm, MotifSpec, showed comparable or marginally improved performance compared to HOMER and Dimont, and markedly improved performance relative to other discriminative motif finding algorithms such as DECOD, DREME and Amadeus when using a fixed search space on mouse and human ChIP-seq data. Since these discriminative motif finders all use similar objective functions, we attribute most of the improvement to our use of a PWM motif model rather than *k*-mers or regular expressions. We also analyzed PBM binding data, and here MotifSpec consistently outperformed the platform-specific motif finder Seed-and-Wobble. Seed-and-Wobble is geared specifically towards the analysis of PBM datasets and uses a *k*-mer enrichment score. MotifSpec is able to do better than Seed-and-Wobble, while remaining agnostic to the underlying technology.

We also presented MotifSpec's novel dynamic expression clustering mode, where we used MotifSpec to search for proximal cis-regulatory elements in yeast and *C. elegans* using sequence and expression similarity on an equal footing. Using this approach on worm, we found a non-redundant list of 87 motifs, which are highly specific for gene target co-expression and functional enrichment. These putative cis-regulatory elements are prime candidates for experimental verification. Only 49 of these elements were found in our earlier investigation of worm expression data [25], and only 12 are in the list of 61 motifs found by the FIRE algorithm [44]. Two elements that we have experimentally validated (data not shown) have been shown to be functional in the worm germline [45].

Recently, significant progress has been made detecting TF-binding sequence signals in more complex mammalian enhancers using SVMs [32,46–48]. In these approaches, all *k*-mers of a given length receive a weight quantifying their importance defining the enhancer set. Because this *k*-mer list is typically very long, it can be difficult to interpret. Because the *k*-mer weight list has a continuous score and is logically identical to PBM data in structure, where the SVM weight is analogous to the PBM enrichment score, we anticipate that MotifSpec may be useful in summarizing important TF binding sites in weights from an SVM trained on mammalian enhancers.

In summary, we have shown that our discriminative motif finder with dynamic search spaces is comparable to or marginally outperforms all of the best existing motif discovery tools, and should improve the extraction of biologically meaningful regulatory elements from the large amounts of ChIP-seq and RNA-seq expression data being generated by high throughput sequencing technologies.

Supporting Information

S1 Fig. Q-Q plots showing fit of PBM data to log-normal distribution. Q-Q plots for two experiments, Bas1 and Pho4-9, are shown.
(PDF)

S2 Fig. Effect of background model on MotifSpec ChIP-seq results. The graphs show the impact on auROC for three datasets of the order of the background model. There was very little change in auROC with the use of higher-order background models.
(PDF)

S3 Fig. MotifSpec uses multiple worker processes to parallelize the motif search process. The worker threads output the motifs found, which are collected by an archiver process that creates a non-redundant archive of motifs. The motif archive is fed back into the worker processes for early termination of searches that are similar to a motif that has already been found.
(PDF)

S4 Fig. Precision recall curve for CTCF. The AUC values are shown in the bottom left corner.
(PDF)

S5 Fig. Precision recall curve for NRSF. The AUC values are shown in the bottom left corner.
(PDF)

S6 Fig. Precision recall curve for ER. The AUC values are shown in the bottom left corner.
(PDF)

S7 Fig. MotifSpec motifs predict PBM binding data better regardless of p-value threshold. The bar charts shows the improvement in the area under the receiver-operator characteristic (ROC) curve, and the top motif found by MotifSpec performs better than the Seed-and-Wobble motif regardless of the p-value threshold used to define the positive set of bound probes. The top chart shows the auROC improvement with a threshold of 0.01 and the bottom chart is with threshold 0.05.
(PDF)

S1 Table. Motifs found in C. elegans by dynamic expression clustering.
(PDF)

Acknowledgments

We would like to thank our lab members Dongwon Lee, Mahmoud Ghandi, Donavan Cheng, and Jun Kyu Rhee for useful discussions and constructive feedback. We would also like to thank Dongwon Lee for his assistance in generating the negative sequence datasets. M. Beer was supported by the Searle Scholars program and NIH grant HG007348.

Author Contributions

Conceived and designed the experiments: RK MB. Performed the experiments: RK MB. Analyzed the data: RK MB. Wrote the paper: RK MB.

References

1. Blat Y, Kleckner N. Cohesins Bind to Preferential Sites along Yeast Chromosome III, with Differential Regulation along Arms versus the Centric Region. Cell. 1999; 98: 249–259. doi: 10.1016/S0092-8674(00)81019-3 PMID: 10428036
2. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science. 2007; 316: 1497–1502. doi: 10.1126/science.1141319 PMID: 17540862
3. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, Bulyk ML. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotech. 2006; 24: 1429–1435. doi: 10.1038/nbt1246
4. Hughes JD, Estep PW, Tavazoie S, Church GM. Computational identification of Cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J Mol Biol. 2000; 296: 1205–1214. doi: 10.1006/jmbi.2000.3519 PMID: 10698627
5. Zhu C, Byers K, McCord R, Shi Z, Berger M, Newburger D, et al. High-resolution DNA binding specificity analysis of yeast transcription factors. Genome Res. 2009; doi: 10.1101/gr.090233.108
6. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, et al. Integrative Analysis of the Caenorhabditis elegans Genome by the modENCODE Project. Science. 2010; 330: 1775–1787. doi: 10.1126/science.1196914 PMID: 21177976
7. ENCODE Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489: 57–74. doi: 10.1038/nature11247 PMID: 22955616
8. Nègre N, Brown CD, Ma L, Bristow CA, Miller SW, Wagner U, et al. A cis-regulatory map of the Drosophila genome. Nature. 2011; 471: 527–531. doi: 10.1038/nature09990 PMID: 21430782
9. Newburger DE, Bulyk ML. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. Nucleic Acids Res. 2009; 37: D77–82. doi: 10.1093/nar/gkn660 PMID: 18842628
10. Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, et al. A method to predict the impact of regulatory variants from DNA sequence. Nat Genet. 2015; 47: 955–961. doi: 10.1038/ng.3331 PMID: 26075791
11. Pervouchine DD, Djebali S, Breschi A, Davis CA, Barja PP, Dobin A, et al. Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. Nat Commun. 2015; 6. doi: 10.1038/ncomms6903
12. Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, et al. A comparative encyclopedia of DNA elements in the mouse genome. Nature. 2014; 515: 355–364. doi: 10.1038/nature13992 PMID: 25409824
13. Lin S, Lin Y, Nery JR, Urich MA, Breschi A, Davis CA, et al. Comparison of the transcriptional landscapes between human and mouse tissues. Proc Natl Acad Sci. 2014; 111: 17224–17229. doi: 10.1073/pnas.1413624111 PMID: 25413365
14. Bailey TL, Williams N, Misleh C, Li WW. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res. 2006; 34: W369–373. doi: 10.1093/nar/gkl198 PMID: 16845028
15. Linhart C, Halperin Y, Shamir R. Transcription factor and microRNA motif discovery: The Amadeus platform and a compendium of metazoan target sets. Genome Res. 2008; 18: 1180–1189. doi: 10.1101/gr.076117.108 PMID: 18411406
16. Bailey TL. DREME: Motif Discovery in Transcription Factor ChIP-Seq Data. Bioinformatics. 2011; 27: 1653–1659. doi: 10.1093/bioinformatics/btr261 PMID: 21543442
17. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010; 38: 576–589. doi: 10.1016/j.molcel.2010.05.004 PMID: 20513432

18. 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: e197–e197. doi: 10.1093/nar/gkt831 PMID: 24057214

19. Huggins P, Zhong S, Shiff I, Beckerman R, Laptenko O, Prives C, et al. DECOD: fast and accurate discriminative DNA motif finding. Bioinformatics. 2011; 27: 2361–2367. doi: 10.1093/bioinformatics/btr412 PMID: 21752801

20. Yao Z, Macquarrie KL, Fong AP, Tapscott SJ, Ruzzo WL, Gentleman RC. Discriminative motif analysis of high-throughput dataset. Bioinforma Oxf Engl. 2014; 30: 775–783. doi: 10.1093/bioinformatics/btt615

21. Patel RY, Stormo GD. Discriminative motif optimization based on perceptron training. Bioinformatics. 2014; 30: 941–948. doi: 10.1093/bioinformatics/btt748 PMID: 24369152

22. Fauteux F, Blanchette M, Strömvik MV. Seeder: discriminative seeding DNA motif discovery. Bioinforma Oxf Engl. 2008; 24: 2303–2307. doi: 10.1093/bioinformatics/btn444

23. Redhead E, Bailey TL. Discriminative motif discovery in DNA and protein sequences using the DEME algorithm. BMC Bioinformatics. 2007; 8: 385. doi: 10.1186/1471-2105-8-385 PMID: 17937785

24. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, et al. Transcriptional regulatory code of a eukaryotic genome. Nature. 2004; 431: 99–104. doi: 10.1038/nature02800 PMID: 15343339

25. Beer MA, Tavazoie S. Predicting Gene Expression from Sequence. Cell. 2004; 117: 185–198. PMID: 15084257

26. Roth FP, Hughes JD, Estep PW, Church GM. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotech. 1998; 16: 939–945. doi: 10.1038/nbt1098-939

27. Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee B-K, et al. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res. 2011; 21: 1757–1767. doi: 10.1101/gr.121541.111 PMID: 21750106

28. Welboren W-J, van Driel MA, Janssen-Megens EM, van Heeringen SJ, Sweep FC, Span PN, et al. ChIP-Seq of ERα and RNA polymerase II defines genes differentially responding to ligands. EMBO J. 2009; 28: 1418–1428. doi: 10.1038/emboj.2009.88 PMID: 19339991

29. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, et al. Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. Cell. 2008; 133: 1106–1117. doi: 10.1016/j.cell.2008.04.043 PMID: 18555785

30. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol. 2008; 9: R137. doi: 10.1186/gb-2008-9-9-r137 PMID: 18798982

31. Niu W, Lu ZJ, Zhong M, Sarov M, Murray JI, Brdlik CM, et al. Diverse transcription factor binding features revealed by genome-wide ChIP-seq in C. elegans. Genome Res. 2011; doi: 10.1101/gr.114587.110

32. Lee D, Karchin R, Beer MA. Discriminative prediction of mammalian enhancers from DNA sequence. Genome Res. 2011; 21: 2167–2180. doi: 10.1101/gr.121905.111 PMID: 21875935

33. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, et al. Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. Mol Biol Cell. 1998; 9: 3273–3297. PMID: 9843569

34. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, et al. Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. Mol Biol Cell. 2000; 11: 4241–4257. doi: 10.1091/mbc.11.12.4241 PMID: 11102521

35. Brauer MJ, Huttenhower C, Airoldi EM, Rosenstein R, Matese JC, Gresham D, et al. Coordination of Growth Rate, Cell Cycle, Stress Response, and Metabolic Activity in Yeast. Mol Biol Cell. 2008; 19: 352–367. doi: 10.1091/mbc.E07-08-0779 PMID: 17959824

36. Baugh LR, Hill AA, Slonim DK, Brown EL, Hunter CP. Composition and dynamics of the Caenorhabditis elegans early embryonic transcriptome. Development. 2003; 130: 889–900. doi: 10.1242/dev.00302 PMID: 12538516

37. Hill AA, Hunter CP, Tsung BT, Tucker-Kellogg G, Brown EL. Genomic Analysis of Gene Expression in C. elegans. Science. 2000; 290: 809–812. doi: 10.1126/science.290.5492.809 PMID: 11052945

38. MCKAY SJ, JOHNSEN R, KHATTRA J, ASANO J, BAILLIE DL, CHAN S, et al. Gene Expression Profiling of Cells, Tissues, and Developmental Stages of the Nematode C. elegans. Cold Spring Harb Symp Quant Biol. 2003; 68: 159–170. doi: 10.1101/sqb.2003.68.159 PMID: 15338614

39. Simcha D, Price ND, Geman D. The Limits of De Novo DNA Motif Discovery. PLoS ONE. 2012; 7: e47836. doi: 10.1371/journal.pone.0047836 PMID: 23144830

40. Harris TW, Antoshechkin I, Bieri T, Blasiar D, Chan J, Chen WJ, et al. WormBase: a comprehensive resource for nematode research. Nucleic Acids Res. 2010; 38: D463–D467. doi: 10.1093/nar/gkp952 PMID: 19910365

41. Guha Thakurta D, Schriefer LA, Waterston RH, Stormo GD. Novel transcription regulatory elements in Caenorhabditis elegans muscle genes. Genome Res. 2004; 14: 2457–2468. doi: 10.1101/gr.2961104 PMID: 15574824

42. Schwartz HT, Horvitz HR. The C. elegans protein CEH-30 protects male-specific neurons from apoptosis independently of the Bcl-2 homolog CED-9. Genes Dev. 2007; 21: 3181–3194. doi: 10.1101/gad.1607007 PMID: 18056428

43. Boxem M, van den Heuvel S. C. elegans Class B Synthetic Multivulva Genes Act in G1 Regulation. Curr Biol. 2002; 12: 906–911. doi: 10.1016/S0960-9822(02)00844-8 PMID: 12062054

44. Elemento O, Slonim N, Tavazoie S. A Universal Framework for Regulatory Element Discovery across All Genomes and Data Types. Mol Cell. 2007; 28: 337–350. doi: 10.1016/j.molcel.2007.09.027 PMID: 17964271

45. Linhart C, Halperin Y, Darom A, Kidron S, Broday L, Shamir R. A novel candidate cis-regulatory motif pair in the promoters of germline and oogenesis genes in C. elegans. Genome Res. 2012; 22: 76–83. doi: 10.1101/gr.115626.110 PMID: 21930893

46. Fletez-Brant C, Lee D, McCallion AS, Beer MA. kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. Nucleic Acids Res. 2013; 41: W544–W556. doi: 10.1093/nar/gkt519 PMID: 23771147

47. Ghandi M, Lee D, Mohammad-Noori M, Beer MA. Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features. PLoS Comput Biol. 2014; 10: e1003711. doi: 10.1371/journal.pcbi.1003711 PMID: 25033408

48. Gorkin DU, Lee D, Reed X, Fletez-Brant C, Bessling SL, Loftus SK, et al. Integration of ChIP-seq and machine learning reveals enhancers and a predictive regulatory sequence vocabulary in melanocytes. Genome Res. 2012; 22: 2290–2301. doi: 10.1101/gr.139360.112 PMID: 23019145