bioRxiv preprint doi: https://doi.org/10.1101/133934; this version posted May 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

SArKS: Discovering gene expression regulatory motifs and domains by suffix array kernel smoothing

Dennis C. Wylie[1*], Hans A. Hofmann[1,2,3,4], Boris V. Zemelman[2,4,5,6**]

May 3, 2017

1 Center for Computational Biology and Bioinformatics, University of Texas at Austin
2 Institute for Cellular and Molecular Biology, University of Texas at Austin
3 Department of Integrative Biology, University of Texas at Austin
4 Institute for Neuroscience, University of Texas at Austin
5 Department of Neuroscience, University of Texas at Austin
6 Center for Learning and Memory, University of Texas at Austin

* denniswylie@austin.utexas.edu
** zemelmanb@mail.clm.utexas.edu

Abstract

Experiments designed to assess differential gene expression represent a rich resource for discovering how DNA regulatory sequences influence transcription. Results derived from such experiments are usually quantified as continuous scores, such as fold changes, test statistics and p-values. We present a *de novo* motif discovery algorithm, SArKS, which uses a nonparametric kernel smoothing approach to identify promoter motifs correlated with elevated differential expression scores. SArKS has the capability to smooth over both motif sequence similarity and, in a second pass, over spatial proximity of multiple motifs to identify longer regions enriched in correlative motifs. We applied SArKS to simulated data, illustrating how SArKS can be used to find motifs embedded in random background sequences, and to two published RNA-seq expression data sets, one probing *S. cerevisiae* transcriptional response to anti-fungal agents and the other comparing gene expression profiles among cortical neuron subtypes in *M. musculus*. For both RNA-seq sets we successfully identified motifs whose kernel-smoothed scores were significantly elevated compared to the permutation-estimated background distributions. We found strong similarities between these identified motifs and known, biologically meaningful sequence elements which may help to provide additional context for the results previously published regarding these data sets. Finally, because eukaryotic transcription regulation is highly combinatorial, we also outline how SArKS methods might be extended to discover synergistic motifs.

Introduction

Discrete sequences—of tones, of symbols, or of molecular building blocks—can provide clues to other characteristics of the entities from which they are derived: a phrase in a bird's song can reveal which species it belongs to, the use of an idiomatic expression can pinpoint a speaker's geographic origin, and a specific short string of nucleotide residues can illuminate

bioRxiv preprint doi: https://doi.org/10.1101/133934; this version posted May 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the function of a DNA domain. In these examples, insights are gleaned from informative *motifs*—short subsequences that match some frequently recurring discernible pattern.

Of particular interest to us are motifs in DNA sequences which are informative with regard to patterns of differential gene expression. The identification of such motifs can help to elucidate the manner in which structure (patterns in DNA sequence) mediates function (regulation of gene expression). Because DNA is largely invariant, individual cell properties tend to be determined by their complement of resident proteins. Tight control over protein expression is, therefore, essential for cellular differentiation, identity, and function. While prior efforts have identified sequences that participate in regulating eukaryotic gene expression, the details regarding how and which specific motifs contribute to specific expression profiles are poorly understood. Here we present an analytical approach toward deciphering this fundamental biological puzzle.

Regulation of gene expression is achieved via a number of complementary processes. First, non-coding DNA is replete with short sequences that can bind transcription factors (TFs), proteins whose own expression varies from cell to cell and over the course of development. Second, DNA can be methylated, epigenetically altering the accessibility of regulatory and coding regions to transcriptional machinery. DNA methylation in turn recruits proteins which modify histones and thereby chromatin structure, further impacting accessibility. In this report, we take the latter regulatory strategies into consideration but focus primarily on accessible regions containing TF binding sites.

In the present study, we present a broadly-applicable algorithm for identifying DNA regulatory domains that support differential gene expression. Our strategy is predicated on the following suppositions: (a) gene expression regulatory regimes involve the binding of TFs to their respective sites on non-coding DNA found near, within, or some distance from a gene; (b) TFs act combinatorially to attract and repel transcription machinery; (c) the same TF binding site may appear multiple times within a stretch of DNA, interspersed with other binding sites; (d) the orientation of a TF binding site gains importance closer to the transcription start site (TSS) of the gene; and (e) there is more than one solution: different genes, even those co-expressed within a single cell, may rely on different regulatory mechanisms. As a practical matter, and in accord with these suppositions, we aim to identify TF binding sites in the vicinity of co-expressed genes and scrutinize their arrangement for significant patterns that can then be evaluated experimentally.

Many different methods for the identification of TF binding motifs have been described. Consensus-based methods such as Weeder [1, 2] focus on motifs of length k that occur repeatedly (allowing for small numbers of mismatches) in sequences of interest. Such methods can be efficiently implemented using suffix trees: Weeder in particular follows a suffix tree-based approach originally described in [3] and [4] with an added heuristic restriction on the pattern of allowed mismatches to maintain the efficiency of the recursive search method utilized [1].

Alternately, profile-based methods such as MEME [5–7] (Multiple Expectation-Maximization for Motif Elicitation) fit a profile model (i.e., a matrix composed of the modeled probabilities of each base occurring at each position of a fixed width motif) of a motif to be compared to a background model in order to classify subsequences as either matching the motif or not. MEME fits these profile models using an expectation-maximization (EM) approach, repeatedly computing the degree to which each subsequence fits the profile (E-step) and then recalculating the profile by realigning subsequences based on these fits (M-step).

bioRxiv preprint doi: https://doi.org/10.1101/133934; this version posted May 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Chromatin immunoprecipitation (ChIP)-based techniques (e.g. ChIP-seq) for identifying protein-interacting DNA sequences have led to the application of motif-finding algorithms to larger sequence data sets than was typical during previous decades [8]. Methods like MDScan [9] can take advantage of the ranking of sequences on based on ChIP enrichment to first generate candidate motifs using only the most enriched DNA sequences and then progressively refine these motifs using the full set of detected DNA sequences.

While MDScan uses functional ranking to separate sequences into sets of higher and lower priority to better focus limited analytical resources for motif discovery, it does not attempt to directly compare one set of sequences to the other. In contrast, discriminative motif analysis [10] seeks to identify motifs specifically differentiating one set of sequences (e.g., promoter regions for a set of genes with a given expression pattern) from another (e.g., a set of reference promoter regions). A number of approaches have been applied to this problem, including [11–18]. A popular recent example, DREME [19] (Discriminative Regular Expression Motif Elicitation), employs Fisher's exact test to assess the significance of motif matches in sequences of one set compared to the other, with further refinement of motif profile conducted for satisfactory candidate motifs.

Discriminative approaches incorporate gene-specific information into the motif discovery process—by, e.g., comparing sequences associated with genes with elevated expression in an experimental condition of interest to sequences associated with genes whose expression shows less evidence of elevation—but these methods implicitly assume that genes may be adequately characterized in a binary manner (e.g., elevated vs. not elevated). Given that the information used to establish the contrasting gene sets is often obtained in the form of continuous expression measurements (and derived measures of differential expression such as t-statistics, f-statistics, etc.), with some genes exhibiting extremely divergent expression patterns across conditions while (usually many) others show more modest differences, it may be more useful to develop methods for what might be called "correlative motif discovery" seeking motifs whose presence signals a trend towards higher or lower values of such a continuous measure.

Correlating motifs from sequences (e.g., promoter regions) w_b with associated continuous score values y_b (e.g., measures of differential expression for the genes associated with the promoter regions) would be straightforward if we had some way of quantifying potential motif patterns present within the w_b. The algorithm we propose here (illustrated in Fig 1) builds on this idea by:

1. concatenating all of the sequences w_b into one supersequence x (detailed in Eq (1) below);

2. constructing the suffix array $[s_i]$ of this supersequence (Eq (4)), where i indexes all suffixes of x sorted into lexicographic order;

3. mapping the suffix positions i back to the sequences w_{b_i} from which the beginnings of the associated suffixes are derived (Eq (5)); and finally

4. for each suffix array index i, applying kernel smoothing to locally regress y_{b_j} on suffix position j (Eq (6)): the resulting smoothed scores $\hat{y}_i$ are then proportional to the correlations of the scores y_{b_j} with the local kernel K_{ij} centered at i.

We are thus using the suffix array index i as the aforementioned quantification of the motif pattern corresponding to the first few characters of the suffix of x beginning at character

bioRxiv preprint doi: https://doi.org/10.1101/133934; this version posted May 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

w_b:	D	NA	AN	DRNAY	AY
y_b:	0	4	0	3	0

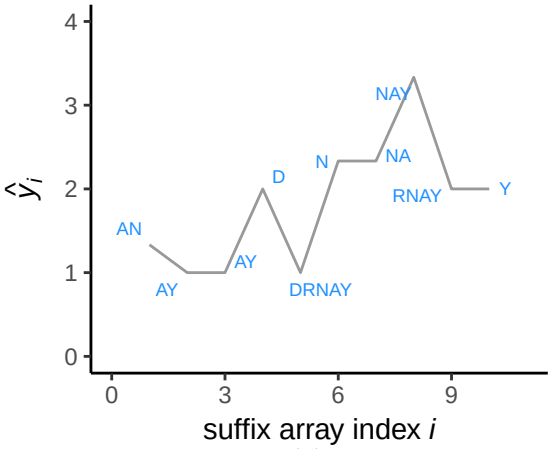
(a)

s	$x[s, \lvert x \rvert)$	b	y_b
0	D NA AN DRNAY AY	0	0
1	NA AN DRNAY AY	1	4
2	A AN DRNAY AY	1	4
3	AN DRNAY AY	2	0
4	N DRNAY AY	2	0
5	DRNAY AY	3	3
6	RNAY AY	3	3
7	NAY AY	3	3
8	AY AY	3	3
9	Y AY	3	3
10	AY	4	0
11	Y	4	0

(b)

i	s_i	$x[s_i, l_{b_i+1})$	b_i	y_{b_i}	$\hat{y}_i$
0	2	A	1	4	nan
1	3	AN	2	0	$4/3$
2	10	AY	4	0	1
3	8	AY	3	3	1
4	0	D	0	0	2
5	5	DRNAY	3	3	1
6	4	N	2	0	$7/3$
7	1	NA	1	4	$7/3$
8	7	NAY	3	3	$10/3$
9	6	RNAY	3	3	2
10	11	Y	4	0	2
11	9	Y	3	3	nan

(c)



(d)

Figure 1: **Overview of SArKS method.** (a) Concatenation of sequences w_b to form string x = DNAAN$DRNAY$AY$ (end-of-sequence character indicated by white space instead of $ for visual clarity). (b) Table of all suffixes of x (part of each suffix following first end-of-sequence character shown in light gray), along with index b of input sequence w_b each suffix derived from and score y_b associated with w_b. (c) Sorted suffix table indicating suffix array index i, suffix array value s_i, suffix (with part following first end-of-sequence character removed), sequence of origin b_i, associated score y_{b_i}, and smoothed score $\hat{y}_i$ generated using smoothing window of size 3 (kernel half-width $\kappa = 1$). (d) Smoothed scores $\hat{y}_i$ plotted against suffix array index i, indicating peak at $i = 8$ corresponding to suffix NAY of input sequence DRNAY. Note that prefix NA of this suffix is longest substring common to the two input sequences w_1 and w_3 with scores $y_b > 0$.

bioRxiv preprint doi: https://doi.org/10.1101/133934; this version posted May 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

s_i. Because i gives the position of a suffix in the lexicographically sorted list of suffixes of the concatenated supersequence x, multiple occurrences of a highly conserved motif—even if they derive from different sequences w—will be consolidated into a run $i, i+1, \ldots, j$ of consecutive index values. Kernel smoothing using a kernel of width on the order $j-i$ thus offers a way to compare the scores $y_{b_i}, y_{b_{i+1}}, \ldots, y_{b_j}$ to the overall score distribution. In this way, **S**uffix **Ar**ray **K**ernel **S**moothing (or SArKS) provides an efficient method for *de novo* discovery of conserved motifs which tend to be found selectively in high-scoring sequences.

We also describe an extension of this method for identification of longer motifs by adding a second round of kernel smoothing applied over the spatial extent of the sequences in order to detect longer regions containing clustered motifs. The use of a nonparametric permutation testing method for computing significance thresholds is then illustrated through the application of SArKS methods to both simulated and real data sets, thus demonstrating (a) the manner in which idealized versions of the motif detection problem may be solved for simulated data and (b) that the algorithm finds plausible candidate patterns with interesting relationships to sequence elements known to have potential regulatory activity when applied to two real gene expression data sets. By implementing a correlational approach to motif discovery, SArKS thus provides a step forward in taking full advantage of the differential expression information offered by RNA-sequencing experiments in the context of motif discovery.

Methods

Motif selection

Given n sequences w_b (also referred to as words) with associated scores y_b, the basic motif selection algorithm defining SArKS consists of:

Concatenation

Concatenate all words w_b (each assumed to end in the line-terminator character \$ lexically prior to all other characters) to form word

$$x = w_0 * w_1 * \ldots * w_{n-1} \tag{1}$$

of length $l_n = |x| = \sum_b |w_b|$. Define also

$$l_b = \sum_{b'<b} |w_{b'}| \tag{2}$$

Thus $x[l_b, l_{b+1}) = w_b$; that is, the substring of the concatenated string starting at position l_b (inclusive) and ending immediately before position l_{b+1} (exclusive) is the sequence w_b (in this paper the first character of a string w is denoted $w[0]$, the second $w[1]$, etc.).

Suffix sorting

Lexically sort suffixes

$$x_s = x[s, |x|) \tag{3}$$

bioRxiv preprint doi: https://doi.org/10.1101/133934; this version posted May 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

into ordered set

$$S = \{x_{s_0}, x_{s_1}, \cdots, x_{s_{l_n-1}}\} \tag{4}$$

thereby defining suffix array $[s_i]$ mapping index i of suffix in S to suffix position s in x (in our software we rely on the Skew algorithm [20] modified to use a difference cover of 7 and implemented in `SeqAn` [21] to efficiently compute the suffix array).

Block marking

Define block array $[b_i]$ by

$$b_i = \max\{b \mid l_b \leq s_i\} \tag{5}$$

mapping index i of suffix in S to block b containing suffix position s_i. The block array then tells us that the character $x[s_i]$ at position s_i in the concatenated string x is derived from $w_{b_i}[s_i - l_{b_i}]$ in the sequence w_{b_i}.

Kernel smoothing

Calculate locally weighted averages

$$\hat{y}_i = \frac{\sum_j K_{ij}\, y_{b_j}}{\sum_j K_{ij}} \tag{6}$$

where the kernel K_{ij} acts as a weighting factor for the contribution of the score y_{b_j} to the smoothing window centered at sorted suffix index i. Loosely speaking, K_{ij} is used to measure how similar (the beginning of) the suffix $x[s_j, |x|)$ is to be considered to (the beginning of) the suffix $x[s_i, |x|)$ in the calculation of a representative score $\hat{y}_i$ averaged over suffixes similar to $x[s_i, |x|)$. As the suffixes have been sorted into lexicographic order, the magnitude of the difference $i - j$ provides some information regarding this similarity: the key idea of the kernel smoothing approach described here is that Eq (6) with K_{ij} defined to be a function of $|i - j|$ may therefore offer a computationally tractable approach for identifying similar substrings (prefixes of suffixes) which tend to occur preferentially in high scoring words w_b.

In this work we use a uniform kernel

$$K_{ij}^{(\kappa)} = \begin{cases} 1 & \text{if } |i-j| \leq \kappa \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

which allows Eq (6) to be computed easily in terms of cumulative sums:

$$\frac{\sum_j K_{ij}^{(\kappa)} y_{b_j}}{\sum_j K_{ij}^{(\kappa)}} = \frac{1}{2\kappa+1} \sum_{j=i-\kappa}^{i+\kappa} y_{b_j} = \frac{1}{2\kappa+1} \left(\sum_{j=1}^{i+\kappa} y_{b_j} - \sum_{j=1}^{i-\kappa-1} y_{b_j} \right) \tag{8}$$

The kernel half-width κ appearing in Eq (7) is an important adjustable parameter in the SArKS methodology controlling the degree of smoothing applied. Increasing κ smooths over more, and hence generally more diverse, suffixes, potentially increasing statistical power at the expense of the resolution of the detected motifs. Recommended guidelines for selecting this parameter are discussed further in Results and discussion.

bioRxiv preprint doi: https://doi.org/10.1101/133934; this version posted May 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

k selection

Set length $\hat{k}_i$ for k-mer associated with suffix array index i by locally averaging the length of suffix sequence identity:

$$\hat{k}_i = \frac{\sum_{j \neq i} K_{ij} \max \{k \leq k_{\max} \mid x[s_j, s_j + k) = x[s_i, s_i + k)\}}{\sum_{j \neq i} K_{ij}} \tag{9}$$

where $k_{\max}$ functions both to increase computational efficiency and to make $\hat{k}_i$ more robust in the presence of a small number of long identical substrings (all results presented here based on $k_{\max} = 12$). Eq (9) is similar to Eq (6) except that: (a) Eq (9) smooths the length (capped at $k_{\max}$) of the longest prefix on which the suffixes $x[s_i, |x|)$ and $x[s_j, |x|)$ agree instead of smoothing the score y_{b_j} as in Eq (6); and (b) Eq (9) omits the central term $i = j$ as it trivially compares suffix the suffix beginning at s_i to itself and is thus uninformative.

Motif selection

Choose score threshold θ and minimum k-mer size $k_{\min}$, thereby defining k-mer set M by

$$M = \left\{ x[s_i, s_i + \lfloor \hat{k}_i \rceil) \mid (\hat{y}_i \geq \theta) \wedge \left(\hat{k}_i \geq k_{\min} \right) \right\} \tag{10}$$

where $\lfloor \hat{k}_i \rceil$ is the nearest integer to $\hat{k}_i$. Strategies for setting the filtering threshold θ based on the permutation testing method described in Permutation testing (and for choosing a reasonable $k_{\min}$) are discussed in Results and discussion.

Limit intra-sequence repeats

One complicating factor in the strategy described in Motif selection is the presence of highly repetitive sequences (common in eukaryotic DNA [22]): if the substring $x[s_i, s_i + rm)$ (assumed to derive wholly from the single word w_{b_i}) consists of $r \gg 1$ repeats of the same m-mer,

$$x[s_i, s_i + rm) = \underbrace{x[s_i, s_i + m)}_{1} * \underbrace{x[s_i, s_i + m)}_{2} * \cdots * \underbrace{x[s_i, s_i + m)}_{r} \tag{11}$$

then it is likely that the sorted suffix array index positions j and k implicitly defined by $s_j = s_i + am$ and $s_k = s_i + bm$ for small $a, b \geq 0$ will be close by, since, assuming without loss of generality that $a < b$,

$$x[s_i + am, s_i + (r - b + a)m) = x[s_i + bm, s_i + rm) \tag{12}$$

showing that the suffixes of x beginning at positions $(s_i + am)$ and $(s_i + bm)$ agree on their first $(r - b)m$ characters. Since all of the positions $s_i + am$ for small a must come from the same word block b_i they must have the same associated score y_{b_i}. If this score y_{b_i} is particularly high, this phenomenon may lead to windows of high $\hat{y}_j$ values centered on j satisfying $s_j = s_i + am$ which result from a very small number of different repeat-containing words (perhaps as few as one if the number of repeats is high enough within a single high-scoring word). An example of a repetitive substring receiving a high smoothed score $\hat{y}_i$ in such manner is discussed in DNA motifs associated with anti-fungal response.

bioRxiv preprint doi: https://doi.org/10.1101/133934; this version posted May 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The distribution of weighted word frequencies

$$f_b^{(i)} = \frac{\sum_j K_{ij}\delta_{b_j b}}{\sum_j K_{ij}} \tag{13}$$

contributing to the window centered at position i of the suffix array table across the full word set W may for these purposes be reasonably summarized by the associated Gini impurity (often used in fitting classification and regression trees [23]):

$$g_i = \sum_b f_b^{(i)} \left(1 - f_b^{(i)}\right) \tag{14}$$

which provides a measure ranging from 0 to $\frac{2\kappa}{2\kappa+1}$ of the degree of uniqueness of the words contributing to the calculation of $\hat{y}_i$. Requiring $g_i \geq g_{\min}$ can thus be used to screen out positions i for which the repeated occurrence of a few high-scoring words in the window centered at i leads to $\hat{y}_i \geq \theta$. Permutation testing further demonstrates that g_i is directly linked to the variation of the smoothed scores $\hat{y}_i$ which would be expected if there were no association between the sequences w_b and the scores y_b, thereby providing the motivation for the use of this particular measure for filtration.

Pruning and extending k-mers

The presence of a k-mer $x[s_i, s_i + k)$ associated with a high smoothed score $\hat{y}_i$ may also result in high smoothed scores $\hat{y}_j$ when $s_j = s_i + m$ if the substring $(k - m)$-mers $x[s_i + m, s_i + k)$ also differentiate higher and lower scoring sequences (if perhaps not as well as the superstring k-mer). The following two steps may be added to the algorithm described in Motif selection in order to reduce the reporting of such substring results in cases where they are present only as part of the full k-mer:

Prune nested k-mers

Cases in which both k-mer $x[s_i, s_i + k)$ and its sub-$(k - m_1 - m_2)$-mer $x[s_i + m_1, s_i + k - m_2)$ (with $m_1 > 0, m_2 \geq 0$) are individually identified can be resolved to report only the longer k-mer: denoting

$$I = \left\{ i \,\middle|\, (\hat{y}_i \geq \theta) \wedge (\hat{k}_i \geq k_{\min}) \wedge (g_i \geq g_{\min}) \right\} \tag{15}$$

remove any index $i \in I$ if there exists $j \in I$ such that the $\lfloor \hat{k}_j \rceil$-mer interval starting at s_j includes all of the $\lfloor \hat{k}_i \rceil$-mer interval starting at s_i, thus retaining only:

$$I' = \left\{ i \in I \,\middle|\, \forall j \in I : \; (s_i \leq s_j) \vee \left(s_i + \lfloor \hat{k}_i \rceil > s_j + \lfloor \hat{k}_j \rceil \right) \right\} \tag{16}$$

This can be done efficiently using an interval tree.

Extend k-mers

For each $i \in I'$, define the duplet

$$\left(z_i^0, z_i^1\right) = \underset{z^0, z^1 \geq 0}{\arg\max} \left\{ z^0 + z^1 \,\middle|\, \exists j \in I' : \; x\left[s_i - z^0, s_i + \lfloor \hat{k}_i \rceil + z^1\right) = x\left[s_j, s_j + \lfloor \hat{k}_j \rceil\right) \right\} \tag{17}$$

bioRxiv preprint doi: https://doi.org/10.1101/133934; this version posted May 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

resolving any ties in the arg max in favor of maximal z^0. Eq (17) picks out the largest super-interval $\left[s_i - z^0, s_i + \lfloor \hat{k}_i \rceil + z^1\right)$ containing the interval $\left[s_i, s_i + \lfloor \hat{k}_i \rceil\right)$ such that the extended $\left(\lfloor \hat{k}_i \rceil + z_i^0 + z_i^1\right)$-mer $x\left[s_i - z^0, s_i + \lfloor \hat{k}_i \rceil + z^1\right)$ is equal to one of the already identified k-mers $\left\{x\left[s_j, s_j + \lfloor \hat{k}_j \rceil\right) \middle| j \in I'\right\}$. Then

$$M' = \left\{x\left[s_i - z_i^0, s_i + \lfloor \hat{k}_i \rceil + z_i^1\right) \middle| i \in I'\right\} \tag{18}$$

defines our pruned motif set.

Spatial smoothing

Existing motif discovery approaches often take into account the tendency of some sequence motifs to exhibit local spatial clustering (thought in some cases to facilitate the cooperative interactions between TFs required for appropriate gene regulation) [24]. Our algorithm can also take advantage of this observation, extending candidate regulatory regions through the application of a second round of kernel-smoothing over the positions within words:

$$\hat{\hat{y}}_{s_i} = \frac{\sum_j L_{s_i t_j} \hat{y}_j}{\sum_t L_{s_i t}} \tag{19}$$

where we here use uniform kernels of the form

$$L_{s_i t_j}^{(\lambda)} = \begin{cases} 1 & \text{if } (0 \leq (t_j - s_i) < \lambda) \wedge (b_i = b_j) \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

(generally with width $\lambda \neq \kappa$) to search for regions of length λ with elevated densities of high-scoring motifs. Note that $\hat{\hat{y}}_{s_i}$ defined by Eq (19) is indexed not by suffix array index i but by suffix array value s_i giving the spatial position s_i in the concatenated word x.

To use such spatial smoothing as an additional basis for motif selection/filtering, it is generally necessary to introduce a second threshold θ_{spatial}, as the doubly-smoothed scores $\hat{\hat{y}}_{s_i}$ will generally be somewhat less dispersed than will be the singly-smoothed $\hat{y}_i$. In this case, formula (21) for the starting motif set M becomes:

$$M = \left\{x[s_i, s_i + \lfloor \hat{k}_i \rceil) \middle| (\hat{y}_i \geq \theta) \wedge (\hat{\hat{y}}_{s_i} \geq \theta_{\text{spatial}}) \wedge (g_i \geq g_{\min}) \wedge (\hat{k}_i \geq k_{\min})\right\} \tag{21}$$

with similar modification to formula (18) for M' then required as well.

Gapped motif detection

While lexical sorting of suffixes assembles occurrences of the same k-mer together into a block of adjacent index positions i, gapped motifs such as

$$u = u_0 * u_{\text{gap}} * u_1 \tag{22}$$

in which there is significant variability in the characters appearing within the internal substring u_{gap} will be scattered into distinct subblocks dispersed within the larger superblock corresponding to their common prefix u_0. By mixing less relevant suffixes in with those

bioRxiv preprint doi: https://doi.org/10.1101/133934; this version posted May 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

corresponding to u within the range of the smoothing kernel, this dispersion can dilute the apparent correlation $\hat{y}_i$ between motif and score.

While the technique described in Spatial smoothing ameliorates this problem to some extent, it does not specifically focus on the important situation where a head motif u_0 is always followed (after the variable u_{gap}) by the same tail motif u_1. We describe here a method for discovering just such gapped motifs by applying first a relatively relaxed threshold θ (which may on its own admit many false positives) and then examining the tail sequences $u_{\text{gap}} * u_1 * \cdots$ following it for evidence of an enriched sequence u_1, pruning away candidate head sequences for which no such corresponding tails can be found.

Defining for any string u:

$$i_{\min}^{(u)} = \min \{i \mid x[s_i, s_i + |u|) \geq u\} \tag{23}$$

$$i_{\max}^{(u)} = \min \{i \mid x[s_i, s_i + |u|) > u\} \tag{24}$$

and noting that $i \in \left[i_{\min}^{(u)},\ i_{\max}^{(u)}\right) \iff x[s_i, s_i + |u|) = u$, we can look for the presence of a particularly common substring u_1 such that the number of occurrences u_1 exactly j positions downstream of an occurrence of u_0 in a sufficiently high-scoring word

$$c_j(u_0, u_1; \theta) = \left| \left\{ i \in \left[i_{\min}^{(u_0)},\ i_{\max}^{(u_0)}\right) \;\middle|\; \hat{y}_i \geq \theta \wedge x\left[i + |u_0| + j,\ i + |u_0| + j + |u_1|\right) = u_1 \right\} \right| \tag{25}$$

is significantly higher than expected. In order to quantify the significance of $c_j(u_0, u_1; \theta)$ some sort of background null model is required; for simplicity we assume homogeneity and independence at different positions i in our examples here, so that according to the null model,

$$c_j(u_0, u_1; \theta) \sim \text{Binom}\left(n(u_0; \theta),\ \prod_{a=j}^{j+|u_1|-1} p_{x[a]} \right) \tag{26}$$

where

$$n(u_0; \theta) = \left| \left\{ i \in \left[i_{\min}^{(u_0)},\ i_{\max}^{(u_0)}\right) \;\middle|\; \hat{y}_i \geq \theta \right\} \right| \tag{27}$$

is the number of occurrences of u_0 in high scoring words (i.e., where $\hat{y}_i \geq \theta$) and $p_{x[a]}$ is the null probability of character $x[a]$. The method here is not constrained to the use of such a naive null model, however; a higher-order Markov null model (as has been demonstrated to improve other motif discovery algorithms [6, 25]) could easily be used instead.

Cluster k-mers by sequence similarity

There are many cases of interest where motifs are not defined by an exact match to a specific k-mer but instead may allow for some variation away from an idealized pattern. Thus the set M defined by Eq (21) is likely to contain many related k-mers which may be more usefully clustered into a few higher-level motif patterns.

Here we adopt a simple edit distance-based criteria to perform this clustering. First we define a diameter $d \geq 0$ controlling how similar motifs must be to cluster together and initialize the (ordered) set of clusters $C = \emptyset$. We then consider the sequences $x[s_i, s_{i+\lfloor \hat{k}_i \rceil})$ in the reverse order of their smoothed scores $\hat{y}_i$ for all suffix indices i surviving all imposed filters, initializing a new cluster "centered" at $x[s_i, s_{i+\lfloor \hat{k}_i \rceil})$ in C if $x[s_i, s_{i+\lfloor \hat{k}_i \rceil})$ is not within

bioRxiv preprint doi: https://doi.org/10.1101/133934; this version posted May 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

d edits of the centers of any existing clusters (e.g., ACGT would initialize a new cluster if $d = 1$ and the only existing cluster was centered on AAG, but would not if a cluster already existed centered at either ACG or AAGT). If, on the other hand, $x[s_i, s_{i+\lfloor \hat{k}_i \rceil})$ is within d edits of the center of one or more existing clusters, it is added to the first such cluster. This clustering strategy has previously been efficiently implemented in the software package `starcode` [26], on which we rely here.

Permutation testing

In order to decide whether the observed correlation between the occurrence of the motifs uncovered by the approach described above and the sequence scores y_b is meaningful, it is useful to have a method for examining results that might be obtained if the sequences w_b and the scores y_b were independent of each other. To this end, the word scores y_b are subjected to permutation π to define

$$y_b^{(\pi)} = y_{\pi(b)} \tag{28}$$

If the permutation π is randomly selected independently of both the sequences w_b and the scores y_b, any true relationships between sequences and scores should be disrupted. This suggests a simple method for assessing the significance of motifs discovered using a given set of parameters (θ, $k_{\min}$, $g_{\min}$, kernel half-width κ, etc.): generate R random permutations π_r and for each permutation select positions i satisfying $\hat{y}_i^{(\pi_r)} \geq \theta$, $\hat{k}_i \geq k_{\min}$, $g_i \geq g_{\min}$, and any other desired criteria (e.g., presence of highly significant tail sequences when searching for gapped motifs as described in Gapped motif detection, or observation of high spatially-smoothed scores $\hat{\hat{y}}_{s_i}$ when the method of Spatial smoothing is employed). In this manner one can estimate the distribution of the number of motifs which would be chosen under a null model in which there is no association between the sequences of the various words w_b and the scores y_b.

This method of significance testing also provides the motivation for the form of Eq (14) in Limit intra-sequence repeats. To demonstrate this, let Π be a random variable representing a random permutation and note that the random variables $y_{\Pi(b)}$ satisfy

$$\mathbb{E}\left[\hat{y}_i^{(\Pi)}\right] = \mathbb{E}\left[\frac{\sum_j K_{ij}\, y_{\Pi(b_j)}}{\sum_j K_{ij}}\right] = \frac{\sum_j K_{ij}\, \mathbb{E}\left[y_{\Pi(b_j)}\right]}{\sum_j K_{ij}} = \bar{y} \tag{29}$$

while, assuming that the number of words $n = |W|$ is large enough that we may approximate $y_{\Pi(b)} \perp\!\!\!\perp y_{\Pi(b')}$ for $b \neq b'$,

$$\mathbb{V}\left[\hat{y}_i^{(\Pi)}\right] = \mathbb{V}\left[\frac{\sum_j K_{ij}\, y_{\Pi(b_j)}}{\sum_j K_{ij}}\right] \approx \sum_b \mathbb{V}\left[f_b^{(i)} y_{\Pi(b)}\right] = \mathbb{V}\left[y_{\Pi(\cdot)}\right] \sum_b \left[f_b^{(i)}\right]^2 \tag{30}$$

where $f_b^{(i)}$ is defined by Eq (13) and for all b

$$\mathbb{V}\left[y_{\Pi(\cdot)}\right] = \mathbb{V}\left[y_{\Pi(b)}\right] = \frac{1}{n}\sum_{b'} (y_{b'} - \bar{y})^2 \tag{31}$$

Eq (30) then tells us that

$$\mathbb{V}\left[\hat{y}_i^{(\Pi)}\right] \propto \left[f_b^{(i)}\right]^2 = 1 - g_i \tag{32}$$

bioRxiv preprint doi: https://doi.org/10.1101/133934; this version posted May 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

where the Gini impurity g_i is defined by Eq (14). Thus smaller values of g_i imply higher variance $\mathbb{V}\left[y_b^{(\Pi)}\right]$ of the window-smoothed scores obtained under random permutation Π (with mean unchanged). This increased variance will lead to the requirement of larger cutoff values θ for reporting motifs discovered in the unpermuted data with a given degree of confidence unless positions i with $g_i < g_{\min}$ are filtered out as described in Limit intra-sequence repeats.

RNA-seq expression analysis

In order to test SArKS, we selected two RNA-seq data sets from Gene Expression Omnibus database [27] (`https://www.ncbi.nlm.nih.gov/geo/`): GSE80357, from *Saccharomyces cerevisiae* (strain 288c), and GSE63137, from *Mus musculus* neocortical neurons [28]. Strain 288c data was obtained following exposure of yeast cells to two different anti-fungal agents. The GSE63137 data set contains detailed transcriptomic and epigenetic information from three distinct non-overlapping classes of pooled neocortical neurons: principal excitatory neurons, parvalbumin (PV)-positive GABAergic neurons, and vasoactive intestinal peptide (VIP)-positive GABAergic neurons.

For the yeast data set GSE80357, we based the sequence scores y_b on the provided gene-level `edgeR` differential expression results:

$$y_b = \begin{cases} \log \Lambda_b & \text{if } \Lambda_b > 0 \\ 0 & \text{otherwise} \end{cases} \tag{33}$$

(where Λ_b is the `edgeR` likelihood ratio statistic for gene b provided in the analysis results for GSE80357).

Because the position of the first used exon often provides information on which TSS is used—and hence on what DNA region defines the applicable promoter—in multicellular eukaryotes, we reanalyzed the GSE63137 RNA-seq data at the transcript level, using `kallisto` [29] to quantify and normalize transcript level expression against Ensembl mouse cDNA reference GRCm38 [30]. Both mean and variance filters were applied (retaining only transcripts for which at least 100 pseudocounts were obtained when summed across all samples, whose mean normalized expression met or exceeded the median of the transcript mean normalized expression levels, and whose normalized expression variance across full sample set similarly met or exceeded the median such value) to winnow the set of transcripts analyzed [31]. In order to simplify downstream analysis, only the isoform with highest mean expression level across all samples was retained for each detected gene. Finally, as previously analyzed epigenetic information on chromatin accessibility was available from the same study [28], only transcripts for which the transcription start sites were located within ATAC-seq peaks (i.e., were accessible) for all examined neuron classes were retained for analysis. Imposing this condition minimizes the likelihood that epigenetic factors, rather than regulatory sequence characteristics, underlie the variations in gene expression across cell classes.

Differential gene expression was then assessed on normalized expression values via standard Student's t-test comparing data for PV neuron data to excitatory and VIP neuron data, with the resulting t-statistic providing a rough estimate of the gene's enrichment in PV neurons to be used as score y_b for transcript b. To prevent the few very large magnitude t-statistics from unduly influencing motif discovery, we enforced a ceiling of 10 on the magnitude of y_b,

bioRxiv preprint doi: https://doi.org/10.1101/133934; this version posted May 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

so that

$$y_b = \begin{cases} -10 & \text{if } t_b \leq -10 \\ t_b & \text{if } -10 < t_b < 10 \\ 10 & \text{if } t_b \geq 10 \end{cases} \tag{34}$$

Results and discussion

Illustration using simulated data

To illustrate the method, we first applied it to a simple simulated data set in which 30 random sequences w_b were generated with each letter $w_b[s]$ drawn independently from a Unif{A,C,G,T} distribution; to the last 10 sequences (i.e., those w_b with $b \geq 20$) we then embedded the k-mer motif CATACTGAGA ($k = 10$) by choosing a position s_b (independently for each sequence w_b) from Unif$\{0, \ldots, |w_b| - k\}$ and replacing $w_b[s_b, s_b + k)$ by the desired k-mer sequence. Scores were assigned to the sequences according to whether the motif had been embedded:

$$y_b = \begin{cases} 0 & \text{if } b \in [0, 20) \\ 1 & \text{if } b \in [20, 30) \end{cases} \tag{35}$$

The kernel half-width $\kappa = 4$ was chosen for this simulation in order to obtain smoothing windows of approximately the same size as the number of motif-positive sequences, $2 * \kappa + 1 \approx |\{b \mid y_b = 1\}|$. In cases where one might expect that most high-scoring sequences exhibit a single conserved copy of a motif while few low-scoring sequences contain the motif, this may be generalized to provide a reasonable starting point for selection of window size: choose $\kappa \approx \frac{1}{2} |\{b \mid y_b \geq \phi\}|$ where ϕ divides "high-scoring" sequences from "low-scoring" ones.

Fig 2 plots $\hat{y}_i$ as obtained from Eq (6) when the method of Motif selection is followed using a uniform kernel with $\kappa = 4$. The highest peaks in the plot correspond to the positions of various substrings of the embedded motif, and lead to the set M of k-mers defined by the $x[s_i, s_i + \lfloor \hat{k}_i \rceil)$ column of table 1.

Pruning table 1 as described in Pruning and extending k-mers, Eq (16) leaves only the rows for $i \in \{2257, 2258, 2256, 1462, 1458, 1463\}$. Applying Eq (17) then extends the 8-mer ATACTGAG of the rows $i \in \{1462, 1458, 1463\}$ to the full 10-mer, so that, following Eq (18), the final k-mer set $M' = \{\text{CATACTGAGA}\}$.

Permutation testing illustrates the utility of setting a minimum k-mer length $k_{\min}$ and/or a minimum block Gini impurity $g_{\min}$ during motif selection: 190 out of 1000 random permutations generated at least one position $i^{(\pi)}$ for which $\hat{y}^{(\pi)}_{i^{(\pi)}} = 1 \geq \theta$ (where θ was taken to have the maximum possible value of 1), but none of these permutations yield any results if $k_{\min} = 6$ is applied to restrict attention to hexamer or longer motifs. Alternatively, if a relatively stringent minimum Gini impurity $g_{\min} = 0.878$ (selected so that only those i for which $b_{i-\kappa}, b_{i-\kappa+1}, \ldots, b_{i+\kappa}$ are all distinct are retained) is enforced, only 2 of 1000 permutations yield positive results, yielding a 95% CI of (0.024%, 0.72%) for family-wise error rate (FWER).

We repeated the process of generating 30 random sequences, embedding the motif CATACTGAGA into the last 10 of them, and then applying suffix array kernel smoothing to the sequence scores 1000 times. In 999 out of these 1000 iterations, the maximum value

$$\hat{y}_{\max} = \max\{\hat{y}_i \mid g_i \geq g_{\min}\} \tag{36}$$

bioRxiv preprint doi: https://doi.org/10.1101/133934; this version posted May 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

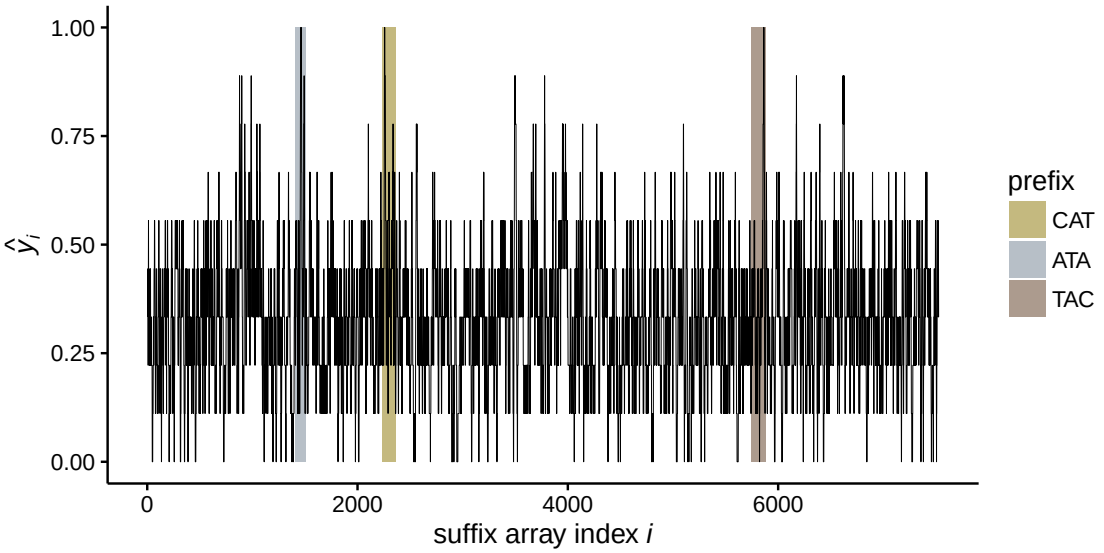


Figure 2: **Locating peaks in kernel-smoothed scores $\hat{y}_i$.** Variation of kernel-smoothed $\hat{y}_i$ (Eq (6), with kernel half-width $\kappa = 4$) along suffix array index i for simulated data set. Gold, silver, and bronze bars indicate positions in suffix array table of suffixes beginning with prefixes CAT, ATA, and TAC, corresponding to first 5 characters of embedded motif CATACTGAGA.

Table 1: **Suffix array positions with $\hat{y}_i \geq \theta$.**

i	s_i	$\hat{y}_i$	$\hat{k}_i$	$x[s_i, s_i + \lfloor \hat{k}_i \rceil)$	b_i	ω_i	g_i
2257	3959	1	10.25	CATACTGAGA	22	194	0.889
2258	4518	1	10.25	CATACTGAGA	25	0	0.889
2256	3544	1	9.62	CATACTGAGA	21	30	0.864
1460	3960	1	9.25	ATACTGAGA	22	195	0.889
1461	4519	1	9.25	ATACTGAGA	25	1	0.889
1459	3545	1	8.75	ATACTGAGA	21	31	0.889
1462	3456	1	8.50	ATACTGAG	20	193	0.864
1458	4442	1	8.25	ATACTGAG	24	175	0.864
5864	3961	1	8.25	TACTGAGA	22	196	0.889
5865	4520	1	8.25	TACTGAGA	25	2	0.889
1463	5595	1	7.88	ATACTGAG	29	73	0.864
5863	3546	1	7.75	TACTGAGA	21	32	0.889
5862	4443	1	7.25	TACTGAG	24	176	0.864
1464	5174	1	7.12	ATACTGA	27	154	0.840
5861	5430	1	6.88	TACTGAG	28	159	0.840
1465	4232	1	6.25	ATACTG	23	216	0.815

Illustration of motif selection process from Motif selection applied to simulated data set with window half-width $\kappa = 4$ and score threshold $\theta = 1$ (here $k_{\min} = g_{\min} = 0$).

bioRxiv preprint doi: https://doi.org/10.1101/133934; this version posted May 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 2: **Unpermuted scores consistently exceed permuted scores only when motif is present.**

$\hat{y}_{\max} - \hat{y}^{(\pi)}_{\max}$	motif (+)	motif (-)
$-2/9$	0	14
$-1/9$	0	165
0	1	650
$1/9$	93	159
$2/9$	790	12
$1/3$	116	0

Distribution of differences between $\hat{y}_{\max}$ obtained by suffix array kernel smoothing using unpermuted sequence scores y_b and $\hat{y}^{(\pi)}_{\max}$ obtained using permuted sequence scores $y_{\pi(b)}$ over 1000 simulations (30 random sequences of 250 characters each) run either with (motif (+) column) or without (motif (-) column) inclusion of motif CATACTGAGA in final 10 sequences.

(with $g_{\min} = 0.878$) calculated using the unpermuted sequence scores exceeded the maximum value

$$\hat{y}^{(\pi)}_{\max} = \max\{\hat{y}^{(\pi)}_i \mid g_i \geq g_{\min}\} \tag{37}$$

obtained using one set of randomly permuted sequence scores per iteration. The full distribution of the differences $\hat{y}_{\max} - \hat{y}^{(\pi)}_{\max}$ is given in the motif (+) column of table 2. Table 2 also contains (motif (-) column) the distribution of $\hat{y}_{\max} - \hat{y}^{(\pi)}_{\max}$ values for 1000 repetitions of an amended version of this process in which the sole modification was to omit the motif embeddings: in this case, $\hat{y}_{\max}$ exceeded $\hat{y}^{(\pi)}_{\max}$ in only 171 of the simulations, while $\hat{y}^{(\pi)}_{\max}$ exceeded $\hat{y}_{\max}$ in 179 simulations (with equality between the two holding in the remaining 650 iterations). The symmetry of the distribution of $\hat{y}_{\max} - \hat{y}^{(\pi)}_{\max}$ around 0 in the motif (-) case is to be expected since the scores y_b are independent of the sequences w_b whether permuted or not if no motifs are included.

Gapped motif detection

Following a similar strategy to that laid out in the Illustration using simulated data above, we generated a second simulated data set containing 30 random 250 character control sequences and then embedding a specific motif into the last 10 of them (again defining y_b by Eq (35)) in order to test the gapped-motif detection strategy of Gapped motif detection. In this case, however, the motif was specified as CATA..CTGA, where the periods between CATA and CTGA represent different pairs of bases randomly assigned to each sequence: 1 AG, 1 CA, 3 CG, 1 GA, 1 GT, and 3 TG (the high frequency of G—in 7 out of 10 embeddings—immediately prior to CTGA here resulted purely from random chance).

This is a more challenging motif discovery problem than the one discussed above. We therefore asked whether approach Gapped motif detection enables SArKS to find correlative motifs in small data sets even when there is no single long conserved section.

The assumptions underlying the selection of κ in the initial simulation study (Illustration using simulated data) are not satisfied in the gapped motif simulations, as the head motif

bioRxiv preprint doi: https://doi.org/10.1101/133934; this version posted May 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

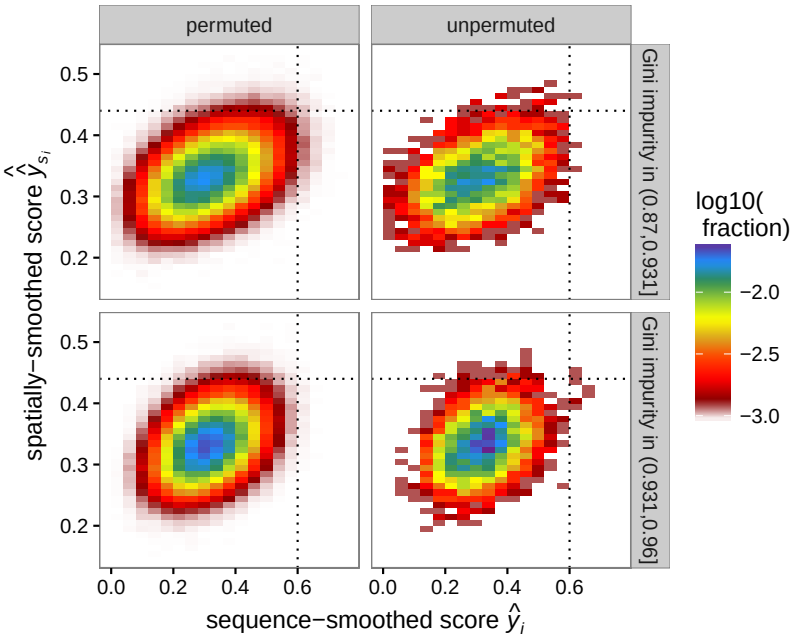


Figure 3: **Joint distribution of smoothed scores and spatially-smoothed scores.** Heatmap depicting binned fractions of suffixes exhibiting different combinations of sequence-smoothed $\hat{y}_i$ (Eq (6)) and spatially-smoothed $\hat{\hat{y}}_{s_i}$ (Eq (19)) values for simulated gapped motif detection problem. Subplots are vertically faceted by Gini impurity g_i (Eq (14)) and horizontally faceted according to whether quantities calculated using permuted scores $y_b^{(\pi)}$ (left) or unpermuted scores y_b (right). Dotted lines correspond the threshold values $\theta = 0.6$ and $\theta_{\text{spatial}} = 0.44$

4-mer $u_0 = \text{CATA}$ is not sufficiently long to guarantee that it will not be present by random chance. Indeed, given independent equiprobable characters in the concatenated sequence x of length $l = 30 * 250 = 7,500$, we would expect any individual k-mer to appear approximately $4^{-k}l$ times on average; for $k = 4$ and $l = 7,500$ this yields an expectation of ≈ 29 random occurrences (in addition to embedded occurrences), or about one per sequence—including the first 20 sequences into which it was not embedded—in our example.

This introduces a new scale to consider: the expected number of total occurrences of the head motif u_0 (here CATA), which we could approximate in this case by the expected number of random occurrences (29) plus the expected number of embedded occurrences (10) at about 39 (actual number of $u_0 = \text{CATA}$ occurrences in simulated data set was 32). However, while this gives us a rough sense of the ceiling on the window size to which the head u_0 contributes, there will generally be subwindows of the window containing all occurrences of u_0 that are particularly enriched in suffixes of higher-scoring sequences $w_{20}, w_{21}, \ldots, w_{29}$. Thus we targeted window sizes slightly below this expected count (39): the value of $\kappa = 12$, corresponding to a full window size of $2 * \kappa + 1 = 25$ was chosen as the mid-point between $\kappa = 4$ appropriate for motifs very unlikely to occur by chance and the value of $\kappa \approx 20$ corresponding to the expected half-width of the window of all suffixes beginning with u_0.

The spatial length scale $\lambda = 10$ over which to smooth $\hat{\hat{y}}_{s_i}$ for this application is based on the length of the target motif $u_0 * u_{\text{gap}} * u_1 = \text{CATA..CATG}$. It could be argued that a slightly lower value of λ would be more appropriate, since there is little reason to expect that suffixes beginning with the last few characters of u_1 will generate high scores $\hat{y}_i$; results

bioRxiv preprint doi: https://doi.org/10.1101/133934; this version posted May 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 3: **Tail sequences for head CATA in gapped motif simulation.**

j	u_1	$c_j(u_0, u_1; \theta)$	$\log_{10} p$
2	CTGA	10	-18.83
1	GCTGA	7	-16.19
2	CTG	10	-12.86
3	TGA	10	-12.86
1	GCTG	7	-11.99
1	GCT	7	-7.83

Ranked p-values for k-mers u_1 found j positions after $u_0 = \text{CATA}$ $c_j(u_0, u_1; \theta = 0.6) \geq 5$ times in gapped motif simulated set.

similar to those presented for $\lambda = 10$ were obtained using λ as low as 5 (though θ_{spatial} must be set higher for smaller values of λ).

Fig 3 plots the joint distributions of $\hat{y}_i$ and $\hat{\hat{y}}_{s_i}$ for both permuted and unpermuted scores and further split into low- and high-Gini impurity g_i indices, with thresholds $\theta = 0.6$ and $\theta_{\text{spatial}} = 0.44$ indicated by dotted lines. This plot suggests a useful method for selecting θ and θ_{spatial}: repeatedly permute the sequence scores y_b to obtain $y_b^{(\pi)}$ to estimate permuted score distributions; thresholds should then be selected high enough that permuted scores rarely exceed them (after filtering out low Gini impurity suffix indices i; note the tighter distribution of $\hat{y}_i$ values for the permuted, high Gini impurity panel as compared to the corresponding low Gini impurity panel, consistent with Eq (32)). Upon further inspection of the unpermuted, high Gini impurity panel of Fig 3, a few disconnected islands containing some of the high-scoring indices i corresponding to occurrences of CATA..CATG may be observed; note that no such islands appear in the permuted distributions.

Thus having set $\kappa = 12$, $\lambda = 10$, $\theta = 0.6$, $\theta_{\text{spatial}} = 0.44$, and having used the median of the Gini impurities g_i to define $g_{\min} = 0.931$, we followed the methods of Motif selection–Pruning and extending k-mers, modified to incorporate spatial smoothing as described in Spatial smoothing, thereby obtaining $M' = \{\text{CATA}\}$ containing only the embedded head sequence u_0.

Table 3 shows the results of searching for common k-mers ($3 \leq k \leq 6$) u_1 occurring within 10 positions downstream of $u_0 = \text{CATA}$ occurrences ranked using the simple binomial null-model described in Gapped motif detection. The most significant hit found is for the correct motif tail sequence $u_1 = \text{CTGA}$, while the remainder of the table contains various substrings of either CTGA or GTCGA, reflecting the randomly occurring bias favoring G immediately preceding the tail sequence in the simulated data.

For permutation testing of this gapped motif detection problem, a threshold p-value of 10^{-10} was applied to determine if any meaningful downstream hits u_1 were detected: at this threshold, while 202 of 1000 permutations $y_b^{(\pi)}$ resulted in positive detection of a head motif u_0 only 1 of these yielded a positive u_1 motif hit (corresponding to 95% CI (0.0025%, 0.56%) for FWER). These results thus demonstrate that statistical analysis of the composition of trailing tail sequences can complement the basic SArKS approach to facilitate the detection of gapped motifs that might otherwise be missed.

bioRxiv preprint doi: https://doi.org/10.1101/133934; this version posted May 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

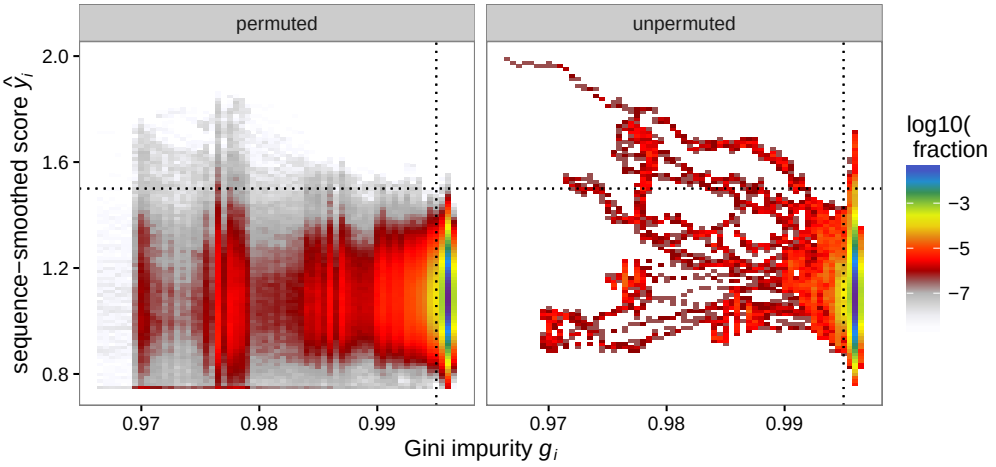


Figure 4: **Joint distribution of smoothed scores and Gini impurities.** Heatmap indicating binned fractions of suffixes exhibiting different combinations of sequence-smoothed $\hat{y}_i$ (Eq (6)) and Gini impurities (Eq (14)) for the yeast data set GSE80357; left panel indicates scores smoothed on permuted data, right panel on unpermuted data. Dotted lines correspond to threshold values $g_{\min} = 0.9950$ and $\theta = 1.5$.

DNA motifs associated with anti-fungal response

We used the methods of Results and discussion to examine potential sequence motifs related to gene expression differences between yeast samples treated with a pair of synergistic anti-fungal agents and a set of matched control specimens as measured in the RNA-seq data set, GEO accession number GSE80357 [32]. The scores y_b for the genes in this data set were derived from the analysis provided in the data set submission as described in RNA-seq expression analysis Eq 33.

The sequences w_b for this application were defined to be the 500 bases immediately upstream (5') of the transcription start site (TSS) of each of 5,436 genes for which `edgeR` [33] analysis results were included in the GSE80357 submission. Using the genome annotations collected in version R64-2-1 of the *S. cerevisiae* gff created by the Saccharomyces Genome Database we calculated that 71.1% of the annotated genes had TSSs were at least 500 bases downstream of the next TSS upstream (median separation between consecutive TSSs calculated to be 888 bases, while mean separation was 1336 bases).

Fig 4 shows the joint distributions of g_i and either $\hat{y}_i$ or $\hat{y}_i^{(\pi)}$ (obtained by smoothing either the true scores y_b or permuted scores $y_b^{(\pi)}$, respectively) as indicated. The propensity for increased variance in the smoothed scores at lower values of g_i underlying Eq (32) can be clearly observed. One consequence of this phenomenon for this data set is that a suffix beginning with a block of 23 consecutive thymine residues simultaneously yields both the highest (unpermuted) $\hat{y}_i$ and the lowest Gini impurity g_i (corresponding to the far-left end of the uppermost red tendril in the right panel of Fig 4); only 53 distinct promoter region sequences w_b contribute to the 251 positions composing the smoothing window centered on this suffix, with 23 of the 251 suffixes derived from a single promoter sequence.

We chose the Gini impurity filter value $g_{\min} = 0.9950$ to satisfy

$$1 - g_{\min} = (1 + \gamma)\left(1 - \operatorname*{median}_i g_i\right) \tag{38}$$

bioRxiv preprint doi: https://doi.org/10.1101/133934; this version posted May 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

with $\gamma = 0.2$, thus removing suffix indices i for which the variance of the permuted smoothed scores $\hat{y}_i^{(\Pi)}$ would be more than approximately 120% of the median value (see discussion leading to Eq (32)); this filter removes only 0.74% of all suffixes from consideration. The threshold $\theta = 1.5$ was determined by examining the distribution of $\hat{y}_{\max}^{(\pi_r)}$ values generated using randomly permuted scores $y_b^{(\pi_r)}$: only 5 of 250 such permutations generated any scores exceeding 1.5 for $\kappa = 125$ and $g_{\min} = 0.9950$ (95% CI (0.65%, 4.6%) for FWER). Using these parameter values and following Motif selection–Pruning and extending k-mers we obtained $M' = \{$TGACTCA, GACTCA, TGACTC, GACTCAT, TGACTAT, ATGACTAA, ATGACTC, TTAGTCA, CCGTACA, AGATAAG, AGATAAGA, GATAAGC, TATATAAG, TATATAAAG$\}$ clustering (setting maximum edit distance $d = 3$) into 3 clusters centered at TGACTCA, CCGTACA, and AGATAAG.

Assessing the similarities of the centers of these 3 high-scoring k-mer clusters to known biological motifs using `tomtom` [34], we found:

TGACTCA similar to binding motif for Yap1p (E-value 0.024)

CCGTACA similar to binding motif for Rap1p (E-value 0.093)

AGATAAG similar to binding motif for Gzf3p (E-value 0.080).

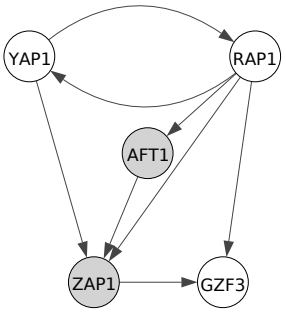


Figure 5: **Regulatory relationships between TFs of interest.** Gray=previously identified, white=corresponds to motif identified here.

While there is little evidence of relevant differential expression for the gene Yap1 (likelihood-ratio (LR)=0.185, $p = 0.67$, $y_{\text{Yap1}} = 0$, $\log_2$-fold-change (logFC)=0.05), the genes for both Rap1 (LR=8.65, $p = 0.0033$, $y_{\text{Rap1}} = 2.16$, logFC=0.31) and Gzf3 (LR=50.5, $p = 1.2e-12$, $y_{\text{Gzf3}} = 3.92$, logFC=0.79) both appear to have elevated expression levels in the simultaneous amphotericin B (AMB) and lactoferrin (LF) treatment group relative to control. Pang et. al. have suggested that the synergistic anti-fungal activity of AMB and LF may involve disruption of oxidative stress response: Yap1 is an essential TF in the normal oxidative stress response [35]. Pang et. al. also discuss the involvement of iron and zinc homeostasis in the synergistic response; Gzf3 has been computationally annotated to Gene Ontology (GO) terms for zinc ion binding and metal ion binding [36, 37]. Furthermore, there is also evidence that the TFs identified with binding sites similar to SArKS identified motifs may regulate or be regulated by TFs previously studied by Pang et. al.: Fig 5 depicts putative regulatory relationships (as found in the YEASTRACT [38] database of documented associations) between these TFs and the two TFs Aft1p and Zap1p previously suggested by [32] as critical actors in the synergistic response of *S. cerevisiae* to the combination of AMB and LF. The distillation of these motifs demonstrates the power of our methodology to uncover candidate sequences that may support differential gene expression.

DNA motifs associated with neuron subtype-specific expression

Finally we applied the SArKS motif discovery methodology to an RNA-seq data set comprising gene expression data for different mouse neocortical neuron subtypes [28]. These authors developed an approach for the purification of genetically defined cell types in mammals and applied it in conjunction with a variety of next-generation sequencing methods—including

bioRxiv preprint doi: https://doi.org/10.1101/133934; this version posted May 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 4: **Expression filters.**

Distinct Transcripts	Count
All	111,669
Detected	73,912
+ Highly Expressed	37,721
+ Highly Varying	29,164
- Duplicate Isoforms	11,857
+ Accessible	6,326

Number of distinct transcript species remaining after sequential application of filters described in RNA-seq expression analysis.

ATAC-seq and MethylC-seq as well as the aforementioned RNA-seq—to investigate epigenetic variation between three different subtypes of neocortical neurons. This data set has many useful features for correlative motif analysis using SArKS beyond the quantification of differential expression, including especially information regarding which regions of the genome are accessible to transcriptional machinery via ATAC-seq. This information is useful not only for filtering the set of genes included in SArKS analysis (as discussed in), but also through the application of ATAC-seq footprint analysis in conjunction with differential methylation analysis that was performed in [28] to infer TF binding at cell-type specific regulatory regions, yielding a set of independently identified TF-binding motifs to which we may compare our own results.

Our initial goal was to identify potential regulatory motifs associated with transcripts enriched in parvalbumin (PV) GABAergic neurons. Because we focused on putative regulatory regions in the vicinity of actively used gene transcription start sites, we quantified expression at the transcript level, filtered transcripts, and determined differential expression as described in RNA-seq expression analysis; table 4 indicates the results of the various transcript filters: 6,326 distinct transcripts, each representing a unique gene, were retained for analysis. For this data set, we conducted three separate SArKS analyses, two focusing upstream (5') of the TSSs for the transcripts of interest and the other downstream (3').

For the selected transcript set, we selected $g_{\min}^{\text{upstream}} = 0.9987$ and $g_{\min}^{\text{downstream}} = 0.9976$ both again using Eq (38) but with the lower value $\gamma = 0.1$ (thus filtering out 9.8% of suffix indices upstream and 5.9% of suffix indices downstream). The lower values here relative to those used for the GSE80357 yeast data set were motivated by the use of longer sequences w_b (appropriate for the less compact mouse genome) increasing the potential for false positive motif signals and thus requiring more stringent thresholds to maintain a high rate of negative results in permutation testing.

Upstream promoter analysis

We first examined upstream sequences w_b for each of the 6,326 remaining transcript species from 3 kb 5' of the TSS to the TSS (the TSSs of 85.5% of mouse genes annotated in Ensembl GRCm38 are separated by greater than 3 kb from the nearest upstream TSS; median separation 23 kb, mean separation 54 kb). Regarding the 979 transcript species whose t-statistic scores $y_b \geq \phi = 2$ for the PV versus other neuron subtypes comparison (see

bioRxiv preprint doi: https://doi.org/10.1101/133934; this version posted May 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

RNA-seq expression analysis for details) as high-scoring, we began with half-window size set at a high-end estimate of $\kappa = 500$ (corresponding to full window size of $2\kappa + 1 = 1001$). In order to select a windowed score threshold θ for the upstream sequence analysis, we generated 250 random permuted score sets $y_b^{(\pi_r)}$ and calculated maximum scores $\hat{y}_{\max}^{(\pi_r)}$ (defined as in Eq (37)) for each: these ranged from 0.424 to 0.560. Based on this distribution, we selected $\theta = 0.55$ as our threshold for this application; 249 out of 250 $\hat{y}_{\max}^{(\pi_r)}$ were less than this threshold (95% CI (0.010%, 2.2%) for FWER). Applying the methods of Motif selection–Pruning and extending k-mers using these sequences as the various w_b and the associated transcript t-statistics as the scores y_b (Eq (34)) with parameters $\kappa = 500$, $\theta = 0.55$, and $g_{\min}^{\text{upstream}} = 0.9987$ resulted in $M'_{\text{upstream}} = \{\text{CCACCTGC, CCACCTGCC}\}$, which clusters into a single motif centered on CCACCTGC for any $d > 0$.

The identified upstream motif sequences CCACCTGC and CCACCTGCC both contain the canonical core recognition E-box sequence CANNTG (more specifically, the E12-box variant CACCTG [39]). Comparison of CCACCTGC with known motifs from the JASPAR database [40] using `tomtom` finds some similarity to TF-binding motifs for SNAI2 (E-value 0.20), MAX (E-value 0.27), SCRT2 (E-value 0.30), SCRT1 (E-value 0.36), and TCF3 (E-value 0.38). 3 of these TFs (SCRT2, SCRT1, and TCF3) were included in the set of genes whose measured expression levels met the minimum mean and variance filters for analysis described in RNA-seq expression analysis; the remaining TFs, SNAI2 and MAX, both met the mean expression criteria but had low expression variance across the 6 analyzed samples. Normalized expression levels of SCRT2 and SCRT1 were elevated in PV neurons relative to excitatory and VIP neurons (t-statistic scores $y_{\text{SCRT2}} = 5.40$ ($p = 0.0057$) and $y_{\text{SCRT1}} = 8.87$ ($p = 0.00089$)), while TCF3 shows little evidence of differential expression between any of the classes of neurons (anova $F_{\text{TCF3}} = 0.59$). Interestingly, the motifs for both SNAI2 and TCF3 were also included in list of TFs identified as possibly regulating cell-type specific expression (for at least one of the 3 cell types studied) using a combination of ATAC-seq footprint analysis and differential methylation analysis in the original study associated with this data set [28].

Use of a smaller smoothing window defined by $\kappa = 250$ for the upstream promoter sequence analysis generated very similar results ($M' = \text{CCACCTGG}$) to those obtained with $\kappa = 500$ but with slightly degraded performance under permutation testing: $\hat{y}_{\max}$ greater than only 243 out of 250 permuted $\hat{y}_{\max}^{(\pi_r)}$ values for $\kappa = 250$ compared to 249 out of 250 permuted scores for $\kappa = 500$. We thus retained the larger $\kappa = 500$ smoothing window here.

Upstream promoter analysis with spatial smoothing

To detect longer regulatory sequences within 3 kb upstream regions of the 6,326 analyzed genes we applied the spatial smoothing method of Spatial smoothing. We retained the same kernel half-window size $\kappa = 500$ and Gini impurity cutoff $g_{\min}^{\text{upstream}} = 0.9987$ and selected the spatial length scale $\lambda = 100$ to target the low end of the enhancer length distribution [41]. Permutation testing then led us to select the combination of $\theta = 0.5$ and $\theta_{\text{spatial}} = 0.25$ (for which there were no positive hits in 250 random permutations (95% CI (0%, 1.5%) FWER)). This resulted in positive hits both for the previously found sequence CCACCTGCC and for 4 closely spaced positions in the $\hat{y}_i$-versus-i plot ($i \in \{8919530, 8919531, 8919548, 891958\}$) corresponding to variations on a lengthy sequence beginning CTGGAACTCACTCTG...; the suffixes corresponding to these 4 peaks were identical in the first 46 nucleotide positions

bioRxiv preprint doi: https://doi.org/10.1101/133934; this version posted May 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and exhibited substantial similarity beyond that.

Because of the high degree of similarity over the longer length of these sequences we bypassed the $\hat{k}_i$ calculations of Eq (9) and instead compared the common first 46 bases of each to known databases, finding an indel-free alignment with 45 of 46 bases perfectly matched with the B1 rodents/Mammalia short interspersed element (SINE) sequence from SINEBase [42]; looking at longer surrounding regions, for each of the 4 peaks we found an alignment to B1 covering at least 132 of the 145 nucleotides in B1 with at least 94% sequence identity. The B1 SINE family consists of retrotransposon-derived sequences which appear repeatedly throughout the mouse genome; recently there have been suggestions that there may be positive selection for the presence of these sequences upstream and in introns of genes with specific functions [43] and that they might also function as enhancers [44].

Downstream promoter analysis

Finally we conducted a third analysis focusing on sequences w_b extending 1 kb downstream (3') of the TSS. Here significant results were obtained using either $\kappa = 500$ or the smaller window $\kappa = 250$; we chose to focus on the $\kappa = 250$ results as they generated longer, potentially more specific, motifs. We chose $\theta = 0.8$ to be higher than all of the $\hat{y}_{\max}^{(\pi_r)}$ resulting from 250 random permutations π_r (the maximum observed $y_{\max}^{(\pi_r)}$ was 0.760; 95% CI for FWER again (0%, 1.5%)) and again applied Motif selection–Pruning and extending k-mers, here setting the Gini impurity cutoff to $g_{\min}^{\text{downstream}} = 0.9976$. The resulting motif set $M'_{\text{downstream}}$ (Eq (18)) contained 7 distinct k-mers: AAGGTCA, ACCTTGG, GACCTTG, GACCTTGG, TGACCTT, TGACCTTG, and TGTCCTTG (with the last of these corresponding to the maximal value of $\hat{y}_i$). Clustering according to Cluster k-mers by sequence similarity ($d = 3$) divides these sequences into two clusters centered at TGACCTTG and AAGGTCA which are clearly reverse complements of the same motif.

Fig 6 shows the distributions of t-statistics for transcript species whose downstream sequences either do or do not contain the highest-scoring octamer TGACCTTG. Comparison of the k-mer sequences with known motifs from the JASPAR database [40] using `tomtom` shows that these k-mers are very similar to the ESRRA/ESRRB/ESRRG binding motifs (e.g., E-value 0.00079 for TGACCTTG match to ESRRA and ESSRB motifs from JASPAR CORE, E-value 0.0046 for match to ESRRG motif). Notably, ESRRA, ESRRB, and ESRRG were all among the previously identified motif set described in [28]. The genes for ESRRA and ESRRG both passed the mean and variance filters employed in RNA-seq expression analysis and both exhibited significantly elevated expression in PV neurons relative to both excitatory and VIP neurons ($y_{\text{ESRRA}} = 3.63$ ($p = 0.022$), $y_{\text{ESRRG}} = 3.34$ ($p = 0.029$)), while ESRRB did not meet the applied mean expression filter (though the low expression levels observed do also indicate elevated expression in PV neurons, $t_{\text{ESRRB}} = 10.2$ ($p = 0.00052$)). TGACCTTG also matched 2 other JASPAR motifs at E-values below 0.1: RORA (E-value 0.0097) and NR5A2 (E-value 0.02). The TF-binding motif for RORA was also in the set of motifs flagged in [28]; neither of the genes RORA nor NR5A2 showed much evidence of elevated expression in PV neurons ($y_{\text{RORA}} = 0.556$, $y_{\text{RXRB}} = 0.825$).

Combining motifs

Fig 7 presents the fractions of analyzed transcript species matching each of the 3 motifs here identified—CCACCTGC and B1 SINE upstream of the TSS and the cluster centered

bioRxiv preprint doi: https://doi.org/10.1101/133934; this version posted May 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

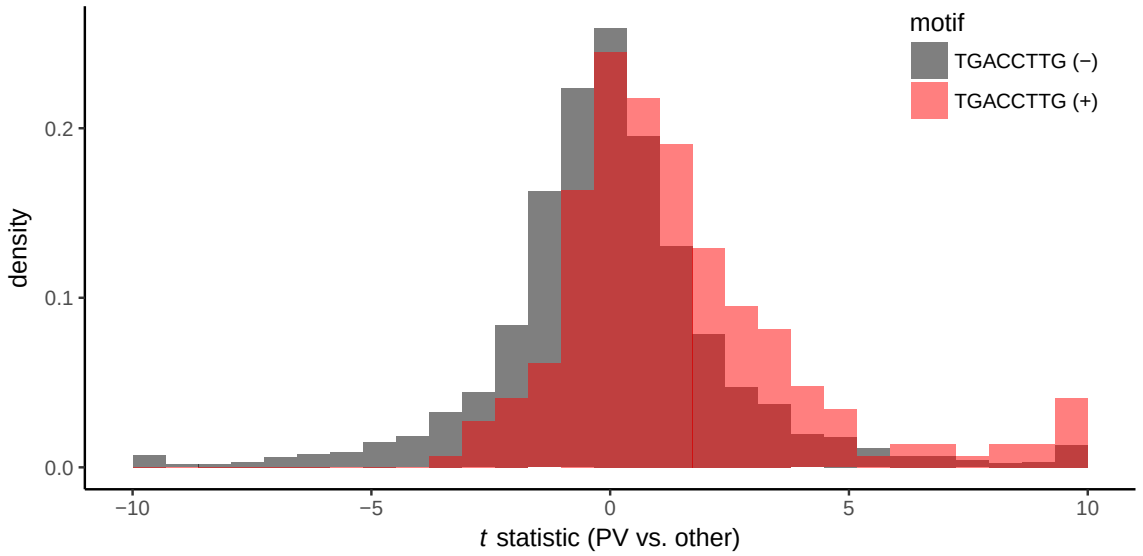


Figure 6: **Sequences containing TGACCTTG tend to be more specifically expressed in PV than sequences not containing TGACCTTG.** Distribution of sequence scores y_b derived from differential expression t-statistics for comparison of PV subtype versus VIP and excitatory subtype neurons (Eq (34)) for transcript species for which k-mer TGACCTTG is either not found (black) or found (red) in first kilobase downstream of transcription start site.

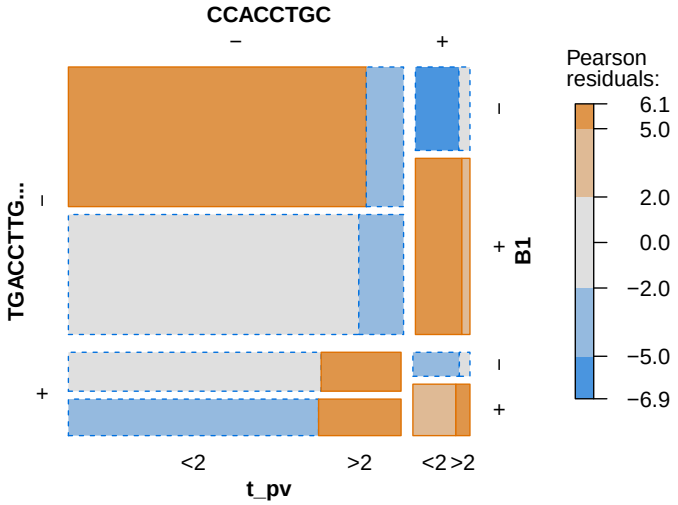


Figure 7: **Combined impact of motifs on differential PV expression.** Mosaic plot indicating co-occurrence trends between motifs identified near transcript TSSs and evidence of difference expression (PV versus other t-statistic binarized with threshold $\phi = 2$). TGACCTTG... represents a match to any of the 7 k-mers composing $M'_{\text{downstream}}$, CCACCTGC represents a match to that specific k-mer, and B1 represents a BLAST match to the mouse B1 repeat sequence identified (alignment length ≥ 70, identical match $\geq 65\%$).

bioRxiv preprint doi: https://doi.org/10.1101/133934; this version posted May 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

on TGACCTTG downstream of the TSS—in a mosaic plot (area of tiles proportional to corresponding fractions. The boxes in the mosaic plot are colored according to whether the observed fraction of sequences containing the motif is above or below the fraction predicted by a null model in which all indicated factors (presence of motif or elevated t-statistic) occur independently: gold=fraction greater than expected under independence, blue=less. Also encoded is the fraction of distinct transcripts with PV-versus-other-subtype t-statistic scores $y_b \geq \phi = 2$. It is apparent that the downstream ESRRA/ESRRB/ESRRG related motif TGACCTTG... has the strongest association with specificity of expression in PV cells, and also that there is a large degree of overlap between the transcript species whose 3 kb upstream regions contain either the E-box CCACCTGC pattern or the B1 sequence pattern (in fact we noted that for some of the highest-scoring B1 matches, a single adenine residue insertion relative to the consensus B1 sequence created a CC**A**CCTGCC match within the B1 region). It is less obvious that the upstream motifs (CCACCTGC or B1) contribute to much increased specificity for those transcript species for which the downstream TGACCTTG... motifs is present. These results suggest that if motifs with more complex combinatorial patterns of association with differential expression are sought it may be useful to take this into account explicitly within the SArKS framework.

Future Directions

Because the regulation of eukaryotic gene expression likely involves interactions among multiple short sequence motifs [45], it is of interest to discover motifs that work together synergistically to confer cell-type specific gene expression profiles. To achieve this objective we need to extend the methods associated with continuous sequence scores introduced in the present study by, e.g., utilizing multivariate kernel regression models such as

$$\hat{y}_{ij} = \frac{\sum\limits_{k,l} K_{ijkl}\, y_{b_k}}{\sum\limits_{k,l} K_{ijkl}} \tag{39}$$

where the 4-index kernel K_{ijkl} might be chosen to satisfy constraints along the lines of

$$K_{ijkl} = \begin{cases} 0 & \text{if } (|i-k| > \kappa_2) \vee (|j-l| > \kappa_2) \\ 0 & \text{if } (b_i \neq b_j) \vee (b_k \neq b_l) \\ 1 & \text{otherwise} \end{cases} \tag{40}$$

That is, i and k must correspond to suffixes with sufficiently similar prefixes (as must j and l), while i and j must come from the same word (as must k and l); see Fig 8.

As discussed in the Introduction, we have made a number of suppositions here regarding the mechanisms by which eukaryotic transcription is regulated, including but not limited to the combinatorial mode of TF action just discussed. A future challenge is to optimally choose the stretch of DNA to be examined relative to the nearby genes: while we have investigated sequences defined solely by proximity to the TSS, it is well known that regulatory elements may lie quite far from their target genes [46]. It would be advantageous to develop more sophisticated approaches to both (a) the identification of genomic regions most likely to contain regulatory elements and (b) the linkage of potential regulatory element-dense

bioRxiv preprint doi: https://doi.org/10.1101/133934; this version posted May 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

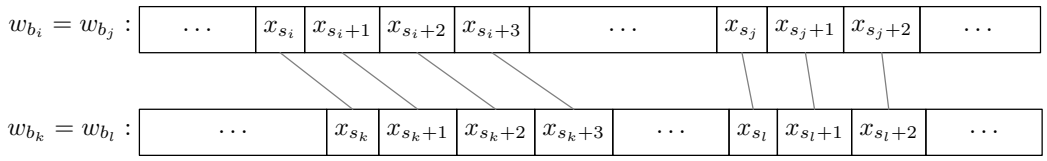


Figure 8: **Potential strategy for identifying synergistic motifs.** Visualization of a quartet (i, j, k, l) satisfying constraints (40): the suffix $x[s_i, |x|)$ originating from sequence w_{b_i} shares a common prefix with the suffix $x[s_k, |x|)$ originating from sequence w_{b_k}; a second pair of suffixes $x[s_j, |x|)$ and $x[s_l, |x|)$ sharing a (different) prefix in common are also found in the sequences $w_{b_j} = w_{b_i}$ and $w_{b_l} = w_{b_k}$, respectively.

regions to governed genes. One place to start may be with information on evolutionary conservation [47] and epigenetic modification [48] near genes of interest.

Finally, while we have tested SArKS on biological sequences, we anticipate uses far afield from this example, including motif discovery in time series data [49], or, by considering node or edge sequences produced by random walks, analysis of complex network structure [50].

Conclusions

We here introduce SArKS as a method for *de novo* correlative motif discovery in order to more fully exploit the results of modern quantitative methods (such as RNA-seq) by avoiding the dichotomization—and consequent loss of information [51]—of sequence scores into discrete groups as required by standard discriminative motif discovery algorithms. SArKS has also been designed with an eye towards minimizing the reliance on specification of specific background sequence models, instead using nonparametric permutation methods [52] to set significance thresholds for motif identification. SArKS is also capable of a second smoothing pass over spatial location of motifs within the sequences in which they are found following the initial smoothing by lexicographic sequence similarity in order to identify longer, potentially interrupted, motifs. Finally, we provide several examples of the usage of SArKS along with detailed analysis of the results thus obtained.

Acknowledgments

This work was supported by BRAIN initiative grant 1U01NS094330 from NINDS and has benefited from discussions with Becca Young, Eric Brenner, and Preeti Mehta.

bioRxiv preprint doi: https://doi.org/10.1101/133934; this version posted May 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

[1] Pavesi G, Mauri G, Pesole G. An algorithm for finding signals of unknown length in DNA sequences. Bioinformatics. 2001;17(suppl 1):S207–S214.

[2] Pavesi G, Mereghetti P, Mauri G, Pesole G. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Research. 2004;32(suppl 2):W199–W203.

[3] Sagot MF. Spelling approximate repeated or common motifs using a suffix tree. In: Latin American Symposium on Theoretical Informatics. Springer; 1998. p. 374–390.

[4] Marsan L, Sagot MF. Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. Journal of Computational Biology. 2000;7(3-4):345–362.

[5] Bailey TL, Elkan C. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. Machine Learning. 1995;21(1-2):51–80.

[6] Bailey TL, Williams N, Misleh C, Li WW. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Research. 2006;34(suppl 2):W369–W373.

[7] Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Research. 2009; p. gkp335.

[8] Zambelli F, Pesole G, Pavesi G. Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. Briefings in Bioinformatics. 2012; p. bbs016.

[9] Liu XS, Brutlag DL, Liu JS. An algorithm for finding protein–DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nature Biotechnology. 2002;20(8):835–839.

[10] Sinha S. Discriminative motifs. Journal of Computational Biology. 2003;10(3-4):599–615.

[11] Segal E, Barash Y, Simon I, Friedman N, Koller D. From promoter sequence to expression: a probabilistic framework. In: Proceedings of the sixth annual international conference on Computational Biology. acm; 2002. p. 263–272.

[12] Segal E, Sharan R. A discriminative model for identifying spatial cis-regulatory modules. Journal of Computational Biology. 2005;12(6):822–834.

[13] Redhead E, Bailey TL. Discriminative motif discovery in DNA and protein sequences using the DEME algorithm. BMC Bioinformatics. 2007;8(1):1.

[14] Fauteux F, Blanchette M, Strömvik MV. Seeder: discriminative seeding DNA motif discovery. Bioinformatics. 2008;24(20):2303–2307.

[15] Valen E, Sandelin A, Winther O, Krogh A. Discovery of regulatory elements is improved by a discriminatory approach. PLoS Computational Biology. 2009;5(11):e1000562.

[16] 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. Molecular Cell. 2010;38(4):576-589.

[17] 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(17):2361–2367.

[18] Yao Z, MacQuarrie KL, Fong AP, Tapscott SJ, Ruzzo WL, Gentleman RC. Discriminative motif analysis of high-throughput dataset. Bioinformatics. 2014;30(6):775–783.

[19] Bailey TL. DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics. 2011;27(12):1653–1659.

bioRxiv preprint doi: https://doi.org/10.1101/133934; this version posted May 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[20] Kärkkäinen J, Sanders P. Simple linear work suffix array construction. In: International Colloquium on Automata, Languages, and Programming. Springer; 2003. p. 943–955.

[21] Döring A, Weese D, Rausch T, Reinert K. SeqAn an efficient, generic C++ library for sequence analysis. BMC Bioinformatics. 2008;9(1):11.

[22] Ellegren H. Microsatellites: simple sequences with complex evolution. Nature Reviews Genetics. 2004;5(6):435–445.

[23] Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and Regression Trees. CRC press; 1984.

[24] Wasserman WW, Sandelin A. Applied bioinformatics for the identification of regulatory elements. Nature Reviews Genetics. 2004;5(4):276–287.

[25] Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouze P, et al. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics. 2001;17(12):1113–1122.

[26] Zorita EV, Cuscó P, Filion G. Starcode: sequence clustering based on all-pairs search. Bioinformatics. 2015; p. btv053.

[27] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Research. 2013;41(D1):D991–D995.

[28] Mo A, Mukamel EA, Davis FP, Luo C, Henry GL, Picard S, et al. Epigenomic signatures of neuronal diversity in the mammalian brain. Neuron. 2015;86(6):1369–1384.

[29] Bray N, Pimentel H, Melsted P, Pachter L. Near-optimal RNA-Seq quantification. arXiv preprint arXiv:150502710. 2015.

[30] Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, et al. Ensembl 2016. Nucleic Acids Research. 2015; p. gkv1157.

[31] Bourgon R, Gentleman R, Huber W. Independent filtering increases detection power for high-throughput experiments. Proceedings of the National Academy of Sciences. 2010;107(21):9546–9551.

[32] Pang CNI, Lai YW, Campbell LT, Chen SCA, Carter DA, Wilkins MR. Transcriptome and network analyses in Saccharomyces cerevisiae reveal that amphotericin B and lactoferrin synergy disrupt metal homeostasis and stress response. Scientific Reports. 2017;7:40232.

[33] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–140.

[34] Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome Biology. 2007;8(2):1.

[35] Kuge S, Jones N, Nomoto A. Regulation of yAP-1 nuclear localization in response to oxidative stress. The EMBO Journal. 1997;16(7):1710–1720.

[36] Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, et al. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Research. 2011; p. gkr1029.

[37] Consortium GO, et al. Gene ontology consortium: going forward. Nucleic Acids Research. 2015;43(D1):D1049–D1056.

[38] Teixeira MC, Monteiro PT, Guerreiro JF, Gonçalves JP, Mira NP, dos Santos SC, et al. The YEASTRACT database: an upgraded information system for the analysis of gene and genomic transcription regulation in Saccharomyces cerevisiae. Nucleic Acids Research. 2013; p. gkt1015.

bioRxiv preprint doi: https://doi.org/10.1101/133934; this version posted May 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[39] Bouard C, Terreux R, Honorat M, Manship B, Ansieau S, Vigneron AM, et al. Deciphering the molecular mechanisms underlying the binding of the TWIST1/E12 complex to regulatory E-box sequences. Nucleic Acids Research. 2016; p. gkw334.

[40] Mathelier A, Fornes O, Arenillas DJ, Chen Cy, Denay G, Lee J, et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Research. 2015; p. gkv1176.

[41] Loots GG. Genomic identification of regulatory elements by evolutionary sequence comparison and functional analysis. Advances in Genetics. 2008;61:269–293.

[42] Vassetzky NS, Kramerov DA. SINEBase: a database and tool for SINE analysis. Nucleic Acids Research. 2013;41(D1):D83–D89.

[43] Tsirigos A, Rigoutsos I. Alu and B1 repeats have been selectively retained in the upstream and intronic regions of genes of specific functional classes. PLoS Computational Biology. 2009;5(12):e1000610.

[44] Elbarbary RA, Lucas BA, Maquat LE. Retrotransposons as regulators of gene expression. Science. 2016;351(6274):aac7247.

[45] Wunderlich Z, Mirny LA. Different gene regulation strategies revealed by analysis of binding motifs. Trends in Genetics. 2009;25(10):434–440.

[46] Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature. 2012;489(7414):109–113.

[47] Xie X, Lu J, Kulbokas E, Golub TR, Mootha V, Lindblad-Toh K, et al. Systematic discovery of regulatory motifs in human promoters and 3 UTRs by comparison of several mammals. Nature. 2005;434(7031):338–345.

[48] Long HK, Sims D, Heger A, Blackledge NP, Kutter C, Wright ML, et al. Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. Elife. 2013;2:e00348.

[49] Fu Tc. A review on time series data mining. Engineering Applications of Artificial Intelligence. 2011;24(1):164–181.

[50] Masoudi-Nejad A, Schreiber F, Kashani Z. Building blocks of biological networks: a review on major network motif discovery algorithms. IET Systems Biology. 2012;6(5):164–174.

[51] Fedorov V, Mannino F, Zhang R. Consequences of dichotomization. Pharmaceutical Statistics. 2009;8(1):50–61.

[52] Ernst MD. Permutation methods: a basis for exact inference. Statistical Science. 2004;19(4):676–685.