Supplement to:

STREME: **Accurate and versatile sequence motif discovery**

Timothy L. Bailey [1,*]

[1]Department of Pharmacology, University of Nevada, Reno, NV 89557, USA

[*]To whom correspondence should be addressed.

1 Supplemental Methods

1.1 STREME algorithm details

1.1.1 Statistical tests of motif enrichment

When the primary and control sequences have identical length distributions, STREME uses the Fisher exact test to determine if a motif is enriched. Suppose that there are N_p and N_c primary sequences and control sequences, respectively, and there are sites in n_p and n_c of the primary and control sequences, respectively. The Fisher exact test assumes a null model where a site is equally likely likely to be in any sequence. It computes the probability that n_p *or more* primary sequences would contain a site, and n_c *or fewer* control sequences would contain a site, given the values N_p, n_p, N_c and n_c.

When the primary and control sequence sets have different length distributions STREME uses the Binomial test (rather than Fisher's exact test) to estimate the statistical significance of a discovered motif. Suppose that there are N_p primary sequences with average length L_p, and N_c control sequences with average length L_c. Then, on average, the number of possible positions a motif of width w could occupy is approximately $S_p = N_p(L_p - w - 1)$ and $S_c = N_c(L_c - w - 1)$, respectively. So STREME estimates the Bernoulli probability P_b that a site chosen randomly in either of the two sets of sequences actually is in a primary sequence as $P_b = S_p/S_c$. If n_p primary sequences contain a match to a given motif (n_p "successes"), the Binomial test considers the statistical significance of the motif to be the probability that n_p or more primary sequences have matches to the motif. The test estimates the p-value of the motif as the sum of the binomial probabilities of k successes in N_p trials, each with probability of success P_b, where k goes from n_p to N_p,

$$p = \sum_{k=n_p}^{N_p} \binom{N_p}{k} P_b^k (1 - P_b)^{N_p - k}.$$

1.1.2 The generalized suffix tree

STREME builds a generalized suffix tree from a single sequence that is the concatenation of all the input sequences (primary and control), with a unique "separator" character placed between sequences. Every suffix in the concatenated sequence corresponds to a leaf node, and vice-versa. The suffix for a leaf can be gotten by reading the labels on the branches along the path from the root to the leaf. STREME labels each leaf as to whether it corresponds to a suffix starting in primary or control sequence, and with the index of number of that sequence. The path from the root to any node in the tree corresponds to a word that occurs in the concatenated sequence. If the path does not contain the separator character, then the word actually occurs in one or more of the input sequences.

1.1.3 Dataset Preparation

Users of STREME would be very surprised STREME produced different results on two datasets that differ only in the order of the sequences. When STREME randomly moves 10% of the input sequences into a "hold-out" set for use in estimating motif statistical significance, the order of the sequences matters. To ensure that identical sets of input sequences (modulo their order) give identical results, STREME first puts the input sequences into a fixed order by sorting them alphabetically. Next, to avoid any biases that this alphabetical sorting might introduce, STREME then shuffles the order of the sequences using a fixed, user-specifiable random seed. For DNA sequences, STREME considers both strands when sorting them alphabetically, sorting by the alphabetically smaller strand. This ensures that STREME is also insensitive to which DNA strand is given for each sequence.

1.1.4 Seed word evaluation

STREME adds to each node in the tree the counts of how many *sequences* (primary and secondary) contain the word corresponding to the node. It does this using a single pass of depth-first search, recording at each node the primary and control sequence index numbers present in the leaves below it. For "valid" nodes—those corresponding to actual words in the input sequences whose length lies in the desired motif width range—STREME computes a p-value by applying its enrichment test to the counts of primary and control sequences for the node. For each valid node, all its prefixes (of legal length) are called "initial seeds".

STREME further evaluates the best 25 initial seeds of each legal width (best means lowest p-value). Evaluation of a seed begins with converting the word to a PWM using maximum likelihood estimation and a Dirichlet prior with a weight of 0.01. The Dirichlet prior is based on a 0-order Markov model of the control sequences. STREME then performs what we call "score-based matching" using the PWM and depth-first search of the suffix tree. STREME uses the PWM to compute the log-likelihood scores of the (legal-width) words corresponding to each node in the depth-first search. The search is pruned whenever the current node is too deep, or when the log-likelihood score of the word is below 0, indicating that the word is more likely under the background model than under the motif model. (Note that when STREME is using a higher-order Markov model of the control sequences, it uses the higher-order model along with the PWM to compute the correct log likelihood score of each word.) At the end of the search, STREME has recorded a list of nodes that correspond *approximate* matches to the PWM and to every prefix of the PWM. STREME uses these lists to determine the best matching site in each sequence. STREME also uses these lists to determine the counts of primary and control sequences that contain matches to the PWM, and computes the p-values of the counts for the PWM and all of its prefixes. Thus, in one pass of (pruned) depth-first search, STREME has scored 25 initial seeds and their prefixes as potential candidates for further refinement. These words are called "seed words".

1.1.5 Motif refinement

STREME further refines the four seed words of each legal width with the lowest p-values.

Refinement of a seed word begins by repeating the same steps as used above for evaluating initial seeds. STREME sorts the list of the best matching site in each sequence in order of decreasing log likelihood score, and finds the score threshold that yields the best enrichment p-value. STREME then estimates a new PWM (by maximum likelihood estimation) from just the sites above the threshold. These two steps—depth first search, reestimating the PWM—are iterated until the p-value fails to improve.

STREME selects the best final PWM from any of the refined seed words as the best motif, which it reports, and whose sites it then erases.

1.1.6 Motif Erasing

STREME "erases" every site matching the best motif in the primary, control and hold-out sequences (and their reverse complements, if the alphabet is complementable) by converting the sites to the separator character. A site is considered a match if its log likelihood score according to the PWM exceeds the optimal score threshold determined in step 4. All matching sites (not just the best one in each sequence) are erased. To allow a certain amount of overlap between the sites of different motifs, STREME only sets letters in the site to the separator character if the letter's likelihood ratio (according to the PWM) is positive.

1.2 Evaluating motif discovery algorithms using TF ChIP-seq data from K562 cells

1.2.1 Reference motifs

In order to evaluate the accuracy, sensitivity, thoroughness and speed of motif discovery algorithms on ChIP-seq datasets, we identify 40 TF ChIP-seq experiments in K562 cells for which there is a motif derived from high-throughput SELEX data for the same TF, or if not, for a member of the same transcription factor family. We download all 150 ENCODE AWG narrowPeak BED files of TF ChIP-seq experiments in K562 cells from the UCSC server at `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform`. Then, for each narrowPeak file, we create a FASTA file of 500bp sequences centered on each of the peaks. Next, using the CentriMo algorithm [2], we determine which of these ChIP-seq peak files shows central enrichment for high-throughput SELEX-derived motif [1]. For some TFs there are multiple SELEX motifs, and for others there is no motif for the TF itself, but there is a motif for a family member. For each of the 150 TF ChIP-seq experiments we assign the SELEX motif for the TF that is most highly enriched (if one exists) to the ChIP-seq peak file as the "reference" motif for that file. If there is no SELEX motif for the ChIP-ed TF, but there is one for a member of the TF's family, we assign the most enriched family-member motif as the reference. These two rules allowed us to assign reference motifs to 40 of the 150 ENCODE TF ChIP-seq experiments. The names of the 40 ENCODE AWG BED files and the names of their reference motifs (which are available at `http://meme-suite.org/meme-software/Databases/motifs/motif_databases.12.19.tgz` in file `EUKARYOTE/Jolma2013.meme`) are given in Table S1.

1.2.2 Primary and control sequence datasets

For each of the 40 ENCODE TF ChIP-seq experiments, we create a single FASTA file containing the 100bp regions around the peaks by using the `fasta-fetch-centered` tool available in the MEME Suite software package to extract sequences from the *Homo sapiens* genome (hg19), which we download from the UCSC genome browser website (`http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.fa.gz`). We use these files, which contain between 1,233 and 56,058 sequences, as the primary sequence datasets. For each such dataset, we create a control dataset by shuffling each sequence in the primary dataset using the `fasta-shuffle-letters` tool available in the MEME Suite software package. With this tool, the user can specify that the frequencies of words (k-mers) of any size k be preserved. The control dataset sequences are thus 100bp long, and have lower-order statistics matching those of the corresponding primary dataset.

To evaluate the thoroughness of motif discovery algorithms, we first identify a non-redundant subset of the 40 TF ChIP-seq sequence datasets described above, and we then create hybrid datasets containing sequences randomly selected from each of the 21 non-redundant TF ChIP-seq sequence datasets thus identified. To identify the non-redundant datasets, we use the `Tomtom` tool (available in the MEME Suite software package) to compare all 40 *reference motifs* to each other, and then greedily add motifs to the non-redundant reference set as long as their similarity to an already added motif is not too high (reject if Tomtom p-value $\leq 10^{-4}$). This results in a collection of 21 non-redundant reference motifs and their associated ChIP-seq datasets (Table S2). We then create 20 hybrid primary sequence datasets by randomly selecting 100 sequences from each of the 21 non-redundant TF ChIP-seq dataset sequence files. (The sequences are all 100bp long.) Then, for each of the 20 hybrid primary datasets we create a matching control dataset by shuffling letters, as described above. The reason that we create the hybrid primary sequence datasets using only the non-redundant subset of ChIP-seq datasets is that we wish to avoid over-representing any motifs. Since many of the original 40 datasets contain ChIP-seq data for the same TF or for TFs in the same motif family, the reference motifs for some ChIP-seq datasets are very similar to each other. Our redundancy-filtering process eliminates this problem.

Table S1: **ENCODE K562 TF ChIP-seq files and SELEX reference motifs.** ChIP-seq files are available at `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform` with extension `gz`. Motifs are available at `http://meme-suite.org/meme-software/Databases/motifs/motif_databases.12.19.tgz` in file `EUKARYOTE/Jolma2013.meme`.

wgEncodeAwgTfbsSydhK562Atf3UniPk.narrowPeak	ATF4_DBD
wgEncodeAwgTfbsHaibK562Atf3V0416101UniPk.narrowPeak	ATF4_DBD
wgEncodeAwgTfbsSydhK562Atf106325UniPk.narrowPeak	ATF7_DBD
wgEncodeAwgTfbsSydhK562Bhlhe40nb100IggrabUniPk.narrowPeak	BHLHE41_full
wgEncodeAwgTfbsSydhK562CebpbIggrabUniPk.narrowPeak	CEBPB_full
wgEncodeAwgTfbsBroadK562CtcfUniPk.narrowPeak	CTCF_full
wgEncodeAwgTfbsUwK562CtcfUniPk.narrowPeak	CTCF_full
wgEncodeAwgTfbsUtaK562CtcfUniPk.narrowPeak	CTCF_full
wgEncodeAwgTfbsSydhK562E2f4UcdUniPk.narrowPeak	E2F7_DBD
wgEncodeAwgTfbsSydhK562E2f6UcdUniPk.narrowPeak	E2F7_DBD
wgEncodeAwgTfbsHaibK562E2f6V0416102UniPk.narrowPeak	E2F7_DBD
wgEncodeAwgTfbsHaibK562Egr1V0416101UniPk.narrowPeak	EGR1_DBD
wgEncodeAwgTfbsHaibK562Elf1sc631V0416102UniPk.narrowPeak	ELF1_DBD
wgEncodeAwgTfbsSydhK562Elk112771IggrabUniPk.narrowPeak	ELK1_DBD_2
wgEncodeAwgTfbsHaibK562Ets1V0416101UniPk.narrowPeak	ETS1_DBD_1
wgEncodeAwgTfbsHaibK562Gata2sc267Pcr1xUniPk.narrowPeak	GATA3_DBD
wgEncodeAwgTfbsSydhK562Gata1UcdUniPk.narrowPeak	GATA4_DBD
wgEncodeAwgTfbsSydhK562Gata2UcdUniPk.narrowPeak	GATA4_DBD
wgEncodeAwgTfbsSydhK562Irf1Ifna6hUniPk.narrowPeak	IRF8_full
wgEncodeAwgTfbsSydhK562Irf1Ifng30UniPk.narrowPeak	IRF8_full
wgEncodeAwgTfbsSydhK562Irf1Ifna30UniPk.narrowPeak	IRF9_full
wgEncodeAwgTfbsSydhK562MaffIggrabUniPk.narrowPeak	MAFF_DBD
wgEncodeAwgTfbsSydhK562MaxIggrabUniPk.narrowPeak	MAX_DBD_2
wgEncodeAwgTfbsHaibK562MaxV0416102UniPk.narrowPeak	MAX_DBD_2
wgEncodeAwgTfbsHaibK562Mef2aV0416101UniPk.narrowPeak	MEF2A_DBD
wgEncodeAwgTfbsSydhK562Nfe2UniPk.narrowPeak	NFE2_DBD
wgEncodeAwgTfbsHaibK562Nr2f2sc271940V0422111UniPk.narrowPeak	NR2F1_DBD_3
wgEncodeAwgTfbsSydhK562Nrf1IggrabUniPk.narrowPeak	NRF1_full
wgEncodeAwgTfbsSydhK562Rfx5IggrabUniPk.narrowPeak	RFX5_DBD_2
wgEncodeAwgTfbsHaibK562Sp1Pcr1xUniPk.narrowPeak	SP1_DBD
wgEncodeAwgTfbsHaibK562Sp2sc643V0416102UniPk.narrowPeak	SP4_full
wgEncodeAwgTfbsHaibK562SrfV0416101UniPk.narrowPeak	SRF_full
wgEncodeAwgTfbsHaibK562Tead4sc101184V0422111UniPk.narrowPeak	TEAD4_DBD
wgEncodeAwgTfbsSydhK562Usf2IggrabUniPk.narrowPeak	USF1_DBD
wgEncodeAwgTfbsHaibK562Usf1V0416101UniPk.narrowPeak	USF1_DBD
wgEncodeAwgTfbsHaibK562Yy1V0416101UniPk.narrowPeak	YY1_full
wgEncodeAwgTfbsSydhK562Yy1UcdUniPk.narrowPeak	YY1_full
wgEncodeAwgTfbsHaibK562Yy1V0416102UniPk.narrowPeak	YY1_full
wgEncodeAwgTfbsHaibK562Zbtb7asc34508V0416101UniPk.narrowPeak	ZBTB7A_DBD
wgEncodeAwgTfbsSydhK562Znf143IggrabUniPk.narrowPeak	ZNF143_DBD

Table S2: **Non-redundant ENCODE K562 TF ChIP-seq files and SELEX reference motifs.** ChIP-seq files are available at `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform` with extension `gz`. Motifs are available at `http://meme-suite.org/meme-software/Databases/motifs/motif_databases.12.19.tgz` in file `EUKARYOTE/Jolma2013.meme`.

wgEncodeAwgTfbsHaibK562Atf3V0416101UniPk.narrowPeak	ATF7_DBD
wgEncodeAwgTfbsSydhK562Bhlhe40nb100IggrabUniPk.narrowPeak	BHLHE41_full
wgEncodeAwgTfbsSydhK562CebpbIggrabUniPk.narrowPeak	CEBPB_full
wgEncodeAwgTfbsBroadK562CtcfUniPk.narrowPeak	CTCF_full
wgEncodeAwgTfbsHaibK562E2f6V0416102UniPk.narrowPeak	E2F7_DBD
wgEncodeAwgTfbsHaibK562Egr1V0416101UniPk.narrowPeak	EGR1_full
wgEncodeAwgTfbsHaibK562Elf1sc631V0416102UniPk.narrowPeak	ELF1_DBD
wgEncodeAwgTfbsHaibK562Gata2sc267Pcr1xUniPk.narrowPeak	GATA3_DBD
wgEncodeAwgTfbsSydhK562Irf1Ifna6hUniPk.narrowPeak	IRF8_full
wgEncodeAwgTfbsSydhK562MaffIggrabUniPk.narrowPeak	MAFF_DBD
wgEncodeAwgTfbsHaibK562Mef2aV0416101UniPk.narrowPeak	MEF2B_full
wgEncodeAwgTfbsSydhK562Nfe2UniPk.narrowPeak	NFE2_DBD
wgEncodeAwgTfbsHaibK562Nr2f2sc271940V0422111UniPk.narrowPeak	NR2F6_DBD_2
wgEncodeAwgTfbsSydhK562Nrf1IggrabUniPk.narrowPeak	NRF1_full
wgEncodeAwgTfbsSydhK562Rfx5IggrabUniPk.narrowPeak	RFX5_DBD_2
wgEncodeAwgTfbsHaibK562Sp1Pcr1xUniPk.narrowPeak	SP1_DBD
wgEncodeAwgTfbsHaibK562SrfV0416101UniPk.narrowPeak	SRF_full
wgEncodeAwgTfbsHaibK562Tead4sc101184V0422111UniPk.narrowPeak	TEAD4_DBD
wgEncodeAwgTfbsHaibK562Yy1V0416101UniPk.narrowPeak	YY1_full
wgEncodeAwgTfbsHaibK562Zbtb7asc34508V0416101UniPk.narrowPeak	ZBTB7B_full
wgEncodeAwgTfbsSydhK562Znf143IggrabUniPk.narrowPeak	ZNF143_DBD

Table S3: **ENCODE K562 eCLIP files and RNAcompete reference motifs.** eCLIP narrowPeak files are available at `https://www.encodeproject.org/search`. Motifs are available at `http://meme-suite.org/meme-software/Databases/motifs/motif_databases.12.19.tgz` in file `RNA/Ray2013_rbp_Homo_sapiens.meme`.

ENCODE file	*Motif Name*
ENCFF700XBC	FMR1
ENCFF777FHS	FUS
ENCFF310NDD	FXR1
ENCFF363IJZ	FXR2
ENCFF955PCQ	HNRNPA1
ENCFF159HMF	HNRNPC
ENCFF448SCQ	HNRNPK
ENCFF296JHG	HNRNPL
ENCFF977MKB	IGF2BP2
ENCFF124YKO	KHDRBS1
ENCFF046VQV	MATR3
ENCFF956PEQ	PABPC4
ENCFF431FMQ	PCBP1
ENCFF924CZR	PTBP1
ENCFF027CBV	QKI
ENCFF788UFQ	SRSF1
ENCFF449ZHX	SRSF7
ENCFF259NLI	TARDBP
ENCFF298ZAG	TIA1
ENCFF331AFC	U2AF2

Table S4: **ENCODE Gm12878 TF ChIP-seq files and SELEX reference motifs.** ChIP-seq files are available at `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform` with extension `gz`. Motifs are available at `http://meme-suite.org/meme-software/Databases/motifs/motif_databases.12.19.tgz` in file `EUKARYOTE/Jolma2013.meme`.

wgEncodeAwgTfbsHaibGm12878Atf2sc81188V0422111UniPk.narrowPeak	ATF4_DBD
wgEncodeAwgTfbsHaibGm12878Atf3Pcr1xUniPk.narrowPeak	ATF7_DBD
wgEncodeAwgTfbsSydhGm12878Bhlhe40cIggmusUniPk.narrowPeak	BHLHE41_full
wgEncodeAwgTfbsBroadGm12878CtcfUniPk.narrowPeak	CTCF_full
wgEncodeAwgTfbsUwGm12878CtcfUniPk.narrowPeak	CTCF_full
wgEncodeAwgTfbsUtaGm12878CtcfUniPk.narrowPeak	CTCF_full
wgEncodeAwgTfbsSydhGm12878E2f4IggmusUniPk.narrowPeak	E2F7_DBD
wgEncodeAwgTfbsHaibGm12878Ebf1sc137065Pcr1xUniPk.narrowPeak	EBF1_full
wgEncodeAwgTfbsSydhGm12878Ebf1sc137065UniPk.narrowPeak	EBF1_full
wgEncodeAwgTfbsHaibGm12878Egr1Pcr2xUniPk.narrowPeak	EGR1_DBD
wgEncodeAwgTfbsHaibGm12878Elf1sc631V0416101UniPk.narrowPeak	ELF1_DBD
wgEncodeAwgTfbsSydhGm12878Elk112771IggmusUniPk.narrowPeak	ELK1_full_2
wgEncodeAwgTfbsHaibGm12878Ets1Pcr1xUniPk.narrowPeak	ETS1_DBD_1
wgEncodeAwgTfbsHaibGm12878Irf4sc6059Pcr1xUniPk.narrowPeak	IRF4_full
wgEncodeAwgTfbsSydhGm12878MaxIggmusUniPk.narrowPeak	MAX_DBD_2
wgEncodeAwgTfbsHaibGm12878Mef2aPcr1xUniPk.narrowPeak	MEF2A_DBD
wgEncodeAwgTfbsHaibGm12878Mef2csc13268V0416101UniPk.narrowPeak	MEF2B_full
wgEncodeAwgTfbsHaibGm12878Nfatc1sc17834V0422111UniPk.narrowPeak	NFATC1_full_1
wgEncodeAwgTfbsSydhGm12878Nfe2sc22827UniPk.narrowPeak	NFE2_DBD
wgEncodeAwgTfbsSydhGm12878Nrf1IggmusUniPk.narrowPeak	NRF1_full
wgEncodeAwgTfbsHaibGm12878Pax5n19Pcr1xUniPk.narrowPeak	PAX5_DBD
wgEncodeAwgTfbsHaibGm12878Pax5c20Pcr1xUniPk.narrowPeak	PAX5_DBD
wgEncodeAwgTfbsHaibGm12878Pou2f2Pcr1xUniPk.narrowPeak	POU3F4_DBD_1
wgEncodeAwgTfbsSydhGm12878Rfx5200401194IggmusUniPk.narrowPeak	RFX2_DBD_1
wgEncodeAwgTfbsHaibGm12878Runx3sc101553V0422111UniPk.narrowPeak	RUNX3_DBD_2
wgEncodeAwgTfbsHaibGm12878Sp1Pcr1xUniPk.narrowPeak	SP1_DBD
wgEncodeAwgTfbsHaibGm12878SrfPcr2xUniPk.narrowPeak	SRF_full
wgEncodeAwgTfbsHaibGm12878Tcf3Pcr1xUniPk.narrowPeak	TCF3_DBD
wgEncodeAwgTfbsHaibGm12878Tcf12Pcr1xUniPk.narrowPeak	TCF3_DBD
wgEncodeAwgTfbsSydhGm12878Usf2IggmusUniPk.narrowPeak	USF1_DBD
wgEncodeAwgTfbsHaibGm12878Usf1Pcr2xUniPk.narrowPeak	USF1_DBD
wgEncodeAwgTfbsHaibGm12878Yy1sc281Pcr1xUniPk.narrowPeak	YY1_full
wgEncodeAwgTfbsSydhGm12878Yy1UniPk.narrowPeak	YY1_full

Table S5: **Non-redundant ENCODE Gm12878 TF ChIP-seq files and SELEX reference motifs.** ChIP-seq files are available at http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform with extension gz. Motifs are available at http://meme-suite.org/meme-software/Databases/motifs/motif_databases.12.19.tgz in file EUKARYOTE/Jolma2013.meme.

wgEncodeAwgTfbsHaibGm12878Atf2sc81188V0422111UniPk.narrowPeak	ATF7_DBD
wgEncodeAwgTfbsSydhGm12878Bhlhe40cIggmusUniPk.narrowPeak	BHLHE41_full
wgEncodeAwgTfbsBroadGm12878CtcfUniPk.narrowPeak	CTCF_full
wgEncodeAwgTfbsSydhGm12878E2f4IggmusUniPk.narrowPeak	E2F7_DBD
wgEncodeAwgTfbsHaibGm12878Ebf1sc137065Pcr1xUniPk.narrowPeak	EBF1_full
wgEncodeAwgTfbsHaibGm12878Egr1Pcr2xUniPk.narrowPeak	EGR1_full
wgEncodeAwgTfbsHaibGm12878Elf1sc631V0416101UniPk.narrowPeak	ELF1_DBD
wgEncodeAwgTfbsSydhGm12878Elk112771IggmusUniPk.narrowPeak	ELK1_full_2
wgEncodeAwgTfbsHaibGm12878Irf4sc6059Pcr1xUniPk.narrowPeak	IRF4_full
wgEncodeAwgTfbsHaibGm12878Mef2aPcr1xUniPk.narrowPeak	MEF2B_full
wgEncodeAwgTfbsHaibGm12878Nfatc1sc17834V0422111UniPk.narrowPeak	NFATC1_full_1
wgEncodeAwgTfbsSydhGm12878Nfe2sc22827UniPk.narrowPeak	NFE2_DBD
wgEncodeAwgTfbsSydhGm12878Nrf1IggmusUniPk.narrowPeak	NRF1_full
wgEncodeAwgTfbsHaibGm12878Pax5c20Pcr1xUniPk.narrowPeak	PAX5_DBD
wgEncodeAwgTfbsHaibGm12878Pou2f2Pcr1xUniPk.narrowPeak	POU3F4_DBD_1
wgEncodeAwgTfbsSydhGm12878Rfx5200401194IggmusUniPk.narrowPeak	RFX2_DBD_1
wgEncodeAwgTfbsHaibGm12878Runx3sc101553V0422111UniPk.narrowPeak	RUNX3_DBD_2
wgEncodeAwgTfbsHaibGm12878Sp1Pcr1xUniPk.narrowPeak	SP1_DBD
wgEncodeAwgTfbsHaibGm12878SrfPcr2xUniPk.narrowPeak	SRF_full
wgEncodeAwgTfbsHaibGm12878Tcf12Pcr1xUniPk.narrowPeak	TCF3_DBD
wgEncodeAwgTfbsHaibGm12878Yy1sc281Pcr1xUniPk.narrowPeak	YY1_full

1.2.3 Evaluating algorithm performance

To evaluate the accuracy, sensitivity and thoroughness of the motif discovery algorithms, we compare the motifs they discover to all 843 motifs in the Jolma compendium of SELEX-derived motifs [1] using the `Tomtom` tool. Our figure of merit is the best (minimum) Tomtom *p*-value of the similarity between the SELEX reference motif for the ChIP-seq dataset and the motifs reported by the algorithm. For a range of similarity (motif accuracy) thresholds, we count the number of times each algorithm finds a motif at least that similar to the SELEX reference motif for the ChIP-seq dataset. For ease of exposition, we convert the Tomtom *p*-value to our "motif similarity score", which is minus the base-10 logarithm of the *p*-value.

We run the respective motif discovery algorithms with the following options, where the values in angle brackets depend on the particular experiment. (The version numbers of the algorithms are given in parentheses after their names.)

- DREME (5.3.0):
  ```
  dreme -p <primary> -n <control> -m <nmotifs>
  -e 1e6 -oc <outdir>
  ```
  where `<primary>` is the file of primary sequences and `<control>` is the file of control sequences.

- HOMER (4.11):
  ```
  findMotifs.pl <primary> fasta <outdir>
  -fastaBg <control> -basic -len 8,10,12
  -noknown -noweight
  ```
  where `<outdir>` is the directory where HOMER will write its results.

- MEME (5.3.0):
  ```
  meme <primary> -bfile <bfile> -dna -revcomp
  -minw <minw> -maxw <maxw> -nmotifs <nmotifs>
  -text -nostatus
  ```
  where `<bfile>` is a Markov model of the desired order created from the control sequences using the command
  ```
  fasta-get-markov -m <order> -dna <control> >
  <bfile>
  ```

- Peak-motifs (1.256):
  ```
  peak-motifs -i <primary> -ctrl <control> -2str
  -markov <kmer-1> -nmotifs <nmotifs> -outdir
  <outdir> -v 1 -title title -max_seq_len 1000
  -disco oligos,positions -minol 6 -maxol 7
  -no_merge_lengths -origin center -scan_markov 1
  -prefix peak-motifs -noov -img_format png -task
  purge,seqlen,composition,disco,merge_motifs,
  split_motifs,motifs_vs_motifs,timelog,archive,
  synthesis,small_summary,scan
  ```

- ProSampler (20191109):
  ```
  ProSampler -i <primary> -b <control> -m
  <nmotifs> -o <outdir>
  ```

- STREME (5.3.0):
  ```
  streme -p <primary> -n <control> -dna -minw
  <minw> -maxw <maxw> -nmotifs <nmotifs> -kmer
  <kmer> -text
  ```

- Weeder (2.0):
  ```
  weeder2 -f <primary> -O HS -chipseq -maxm
  <nmotifs> -oc <outdir>
  ```

For RNA experiments we make the following changes to the above command lines. For DREME, we add `-rna`. For HOMER, we add `-rna`. For MEME, we replace `-dna -revcom` with `-rna`. For Peak-motifs, we replace `-2str` with `-1str`. For ProSampler, we add `-p 1`. For STREME, we replace `-dna` with `-rna`. For Weeder, we add `-ss`.

1.3 Evaluating motif discovery with CLIP-seq datasets

We follow a similar procedure to the above in identifying and preparing a group of CLIP-seq datasets and reference motifs. The datasets are derived from ENCODE enhanced CLIP-seq (eCLIP) databases from the Yeo lab [6], and the reference motifs come from the RNAcompete database [3]. We manually determine the set of 20 RNA-binding proteins with ENCODE eCLIP datasets from K562 cells that also have (at least one) motif in the RNAcompete compendium for *Homo sapiens*. We then download the eCLIP narrowPeak BED files that combine two eCLIP replicates from the ENCODE website (`https://www.encodeproject.org/search`) for those 20 experiments. From each of the 20 BED files we create a BED file of non-overlapping regions using a custom script. From each non-overlapping BED file we create a FASTA file of sequences using the `fasta-fetch-centered` tool from MEME Suite specifying its `-r` option in order to fetch the exact regions specified in the BED file from the *Homo sapiens* hg19 genome. We use the 20 FASTA files thus created, which contain between 166 and 7,520 sequences with average length 70, as the primary sequence datasets in our evaluation of motif discovery algorithms. The ENCODE accession numbers and reference motifs for the eCLIP datasets are given in Table S3.

We find that using a shuffled version of the primary sequence dataset as the control dataset does not work as well with this eCLIP data and these reference motifs as using a more random control dataset. All the motif discovery algorithms we test here perform better when we construct a single control dataset by consisting of 10,000 sequences randomly chosen from the combined sequences in the 20 primary sequence datasets. This is the control dataset we use here in all eCLIP-based analyses.

We evaluate algorithm performance on the eCLIP datasets analogously to our approach with the ChIP-seq datasets (see previous section).

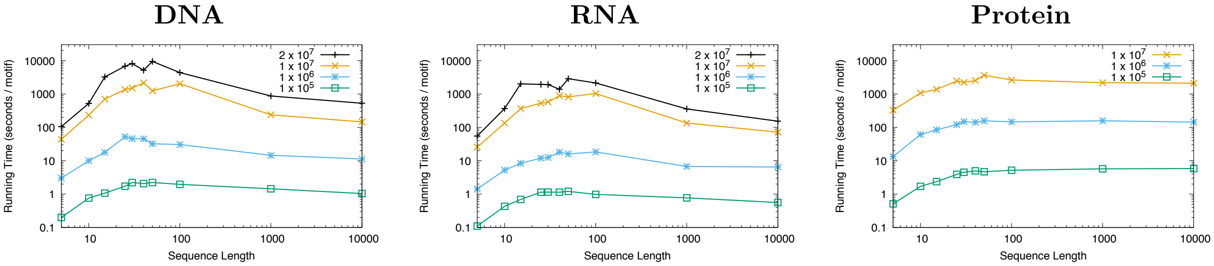


Figure S1: **Running time of STREME on different types, lengths and numbers of sequences.** Each point shows the running time in seconds per motif found (Y) when STREME run with a primary sequence dataset of a given *total size* (color), where all the sequences have the given length (X). The points for a given total size are connected with straight lines for ease of interpretation. The sequences generated using a uniform distribution over the alphabet named above the panel by the MEME Suite `gendb` algorithm. STREME was run on a 3.2 GHz Intel Core i7 processor with 16GB of memory.

1.4 Evaluating STREME with a custom alphabet

To evaluate STREME using a custom alphabet, we create a "corrupted" version of the novel "Moby Dick". We download the complete text of the novel from the Gutenberg Project (`https://gutenberg.org`). To make motif discovery more difficult, we remove all punctuation, spaces and capitalization from the text. We then convert the text to FASTA format, placing each line in the original text file in a single FASTA sequence. To make the problem more similar to a biological one, we then corrupt the text by randomly changing 20% of the letters in each FASTA sequence to a different, randomly-chosen letter. We then run STREME on the resulting FASTA file, along with the alphabet definition file that defines the 26 letter alphabet of the sequences.

1.5 Creating random sequence datasets

To evaluate the accuracy of STREME p-values, and to evaluate speed of STREME on different alphabets and dataset sizes, we use randomly generated sequences. We generate random DNA and protein sequence datasets using the `gendb` tool from the MEME Suite package. This tool allows us to specify the number as well as the minimum and maximum length of generated sequences. We generate RNA sequences by replacing `T` with `U` in DNA sequences generated by `gendb`.

2 Supplemental Results

2.1 Speed of STREME with different alphabets and dataset sizes

We study the effect of the sequence alphabet and the size of the primary dataset using sets of artificially generated (random) sequences. We use datasets containing sequences of a fixed length, and adjust the number of sequences to create datasets containing from 100,000 to 20,000,000 characters (i.e., DNA or RNA bases or protein residues). We let STREME automatically create the control dataset, of equal size as the primary dataset. We allow STREME to report exactly five motifs, and let STREME pick the motif width between 5 and 30 positions. All runs are on a 3.2 GHz Intel Core i7 processor with 16GB of memory.

The results for DNA, RNA and protein sequences are shown in Fig. S1. Overall, the running time of STREME is roughly proportional to $n \log n$, where n is the total primary dataset size in characters. For a given total primary sequence dataset size, running time tends to be maximal when the sequence length is approximately equal to the maximum allowed motif width, regardless of the alphabet size or the total dataset size.

STREME is very fast, discovering motifs in 10,000 DNA sequences of length 100bp in about one minute per motif. With RNA sequences, STREME runs about twice as fast as with DNA since STREME treats RNA as single stranded, making the suffix tree half as large. Because the protein alphabet is larger, causing the suffix tree to have more branches, STREME runs about about 5 times more slowly with protein sequences than with RNA. In general, for single-stranded alphabets, the running time of STREME is roughly proportional to the size of the alphabet, and double-stranded alphabets take twice as long as a single-stranded alphabet of the same size.

2.2 Evaluating the effect of STREME's search parameters `NEVAL` and `NREF`

The effect on the thoroughness of motif discovery with STREME when the number of words evaluated for approximate matches (`NEVAL`), and the number of evaluated words that are converted to motifs and refined (`NREF`) is shown in Fig. S2. The same experimental procedures are used as described for Fig. 3 in the main paper.

It is clear from Fig. S2 that varying `NEVAL` from 5 to 55 (de-

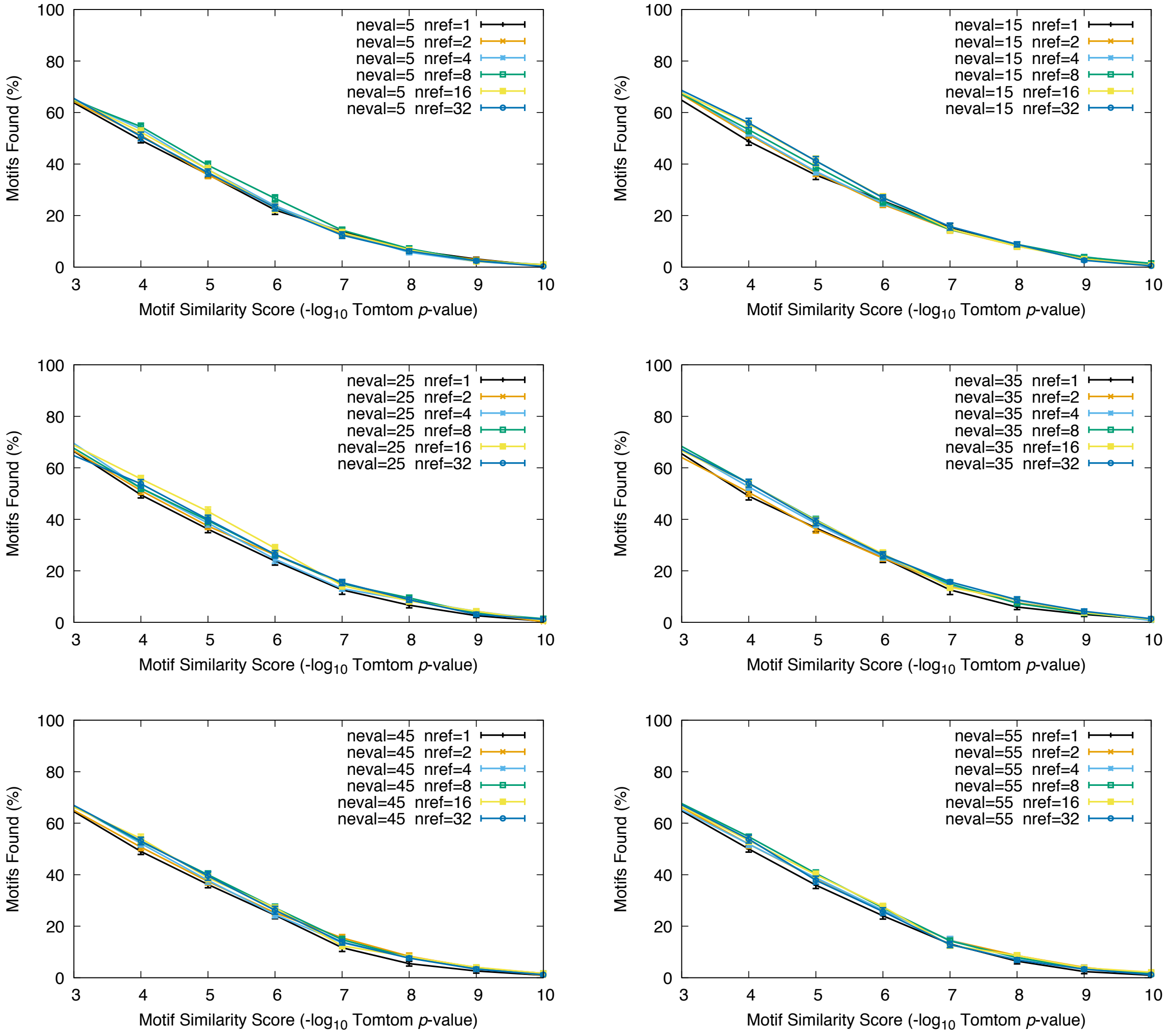


Figure S2: **Thoroughness of STREME as a function of `NEVAL` and `NREF`.** The curves show the percentage of 21 reference motifs (Y) for which the named algorithm finds a motif matching it with given motif similarity score (X) or better, averaged over 20 combined datasets. In each plot, STREME is run with a fixed value of `NEVAL` from 5 to 55, and the curves show results for values of `NREF` from 1 to 32.

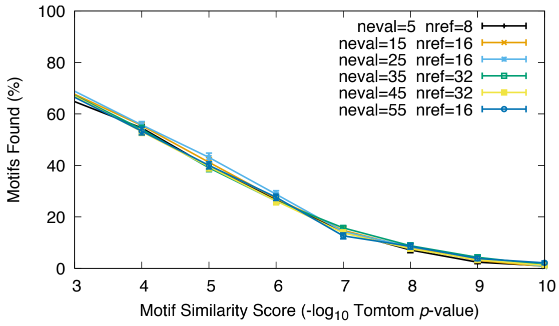


Figure S3: **Thoroughness of STREME with best combinations of `NEVAL` and `NREF`.** The curves show the percentage of 21 reference motifs (Y) for which the named algorithm finds a motif matching it with given motif similarity score (X) or better, averaged over 20 combined datasets. Each curve shows the results for when the best value of `NREF` in the range 1 to 32 was chosen for the given value of `NEVAL`.

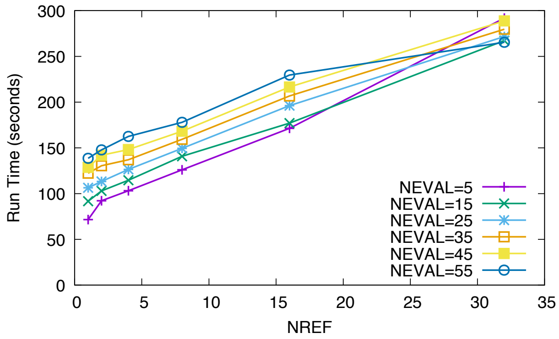


Figure S4: **Running time of STREME as a function of `NEVAL` and `NREF`.** The curves show the running time of STREME in the experiments from Fig. S2. Each curve shows the running time as a function of the value of `NREF` in the range 1 to 32, for a fixed value of `NEVAL`.

fault=25), and `NREF` from 1 to 32 (default=4), has little effect on STREME's ability to discover multiple motifs in a single sequence dataset. There are small differences, and the best value of `NREF` for each value of `NEVAL` is shown in Fig. S3. Balancing the thoroughness in that figure, and the running time (shown in Fig. S4), values of 25 and 16 for `NEVAL` and `NREF` appear close to optimal. These values improve the thoroughness of STREME compared with using its default values for these parameters (25 and 4, respectively) in this experiment, as shown in Fig. S5**c**. However, using these optimal values for `NEVAL` and `NREF` does not improve either the accuracy or sensitivity of STREME, as shown in Fig. S5**c** and Fig. S5**b**. (Those two figures repeat the experiments shown in Fig. 1**a** and Fig. 2, respectively, using the optimal values of `NEVAL` and `NREF` based on thoroughness.)

We conclude that the default values of the parameters `NEVAL` and `NREF` are probably satisfactory, although increasing `NREF` from 4 to 16 might improve STREME's thoroughness. However, this improvement comes at the cost of an increase in running time of STREME of approximately 50% (126 sec. vs 196 sec., see Fig. S4).

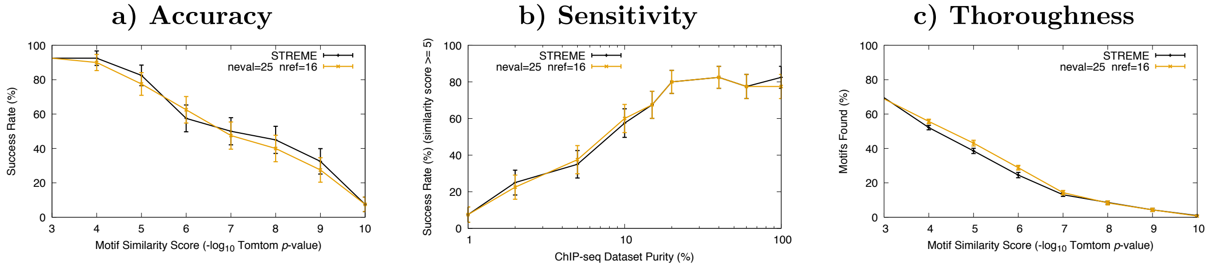
In the next section we evaluate the effect of using the new values of `NEVAL` and `NREF` on an independent group of ChIP-seq datasets from a different cell type–Gm12878. That evaluation does not provide support for changing STREME's default settings for `NEVAL` and `NREF`.

2.3 Evaluating motif discovery algorithms using TF ChIP-seq data from Gm12878 cells

As an independent test of STREME and other motif discovery algorithms, we performed a second evaluation, analogous to that described above in section 1.2, substituting TF ChIP-seq data from Gm12878 cells. The data for the accuracy and sensitivity tests comprise 33 ENCODE AWG BED files and 33 reference motifs (Table S4). The data for the thoroughness tests comprise 21 non-redundant reference motifs and their associated ChIP-seq datasets (Table S5).

The results of this independent analysis using data from Gm12878 cells are consistent with our results using data from K562 cells. STREME has superior accuracy compared to the other motif finders tested (Fig. S6**a**). STREME is more sensitive than all the other tested algorithms, except HOMER, which shows equal sensitivity (Fig. S6**b**). Finally, STREME is more thorough than the tested algorithms (Fig. S6**c**).

We also tested the effect of using the optimal values for the two STREME parameters `NEVAL` and `NREF` that were determined using K562 data. The results of running STREME with its default values (25 and 4, respectively) and the values that were optimal with K562 data (25 and 16, respectively) are shown in Fig. S6**d-f**. Accuracy on Gm12878 data is slightly worse using the K562-optimal parameters, whereas

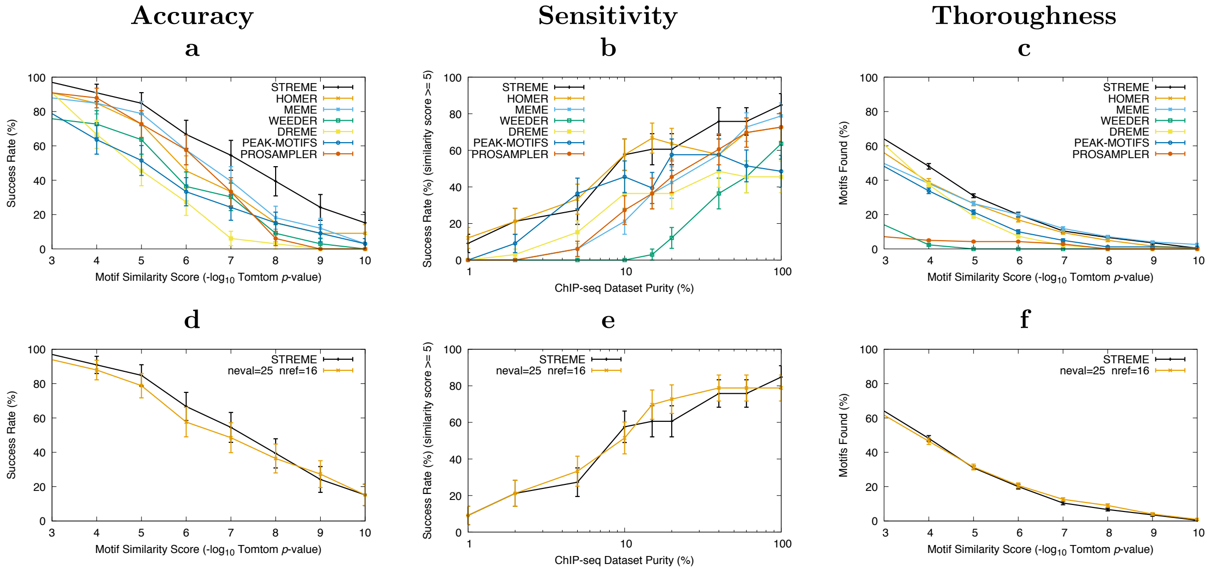


a) Accuracy

b) Sensitivity

c) Thoroughness

Figure S5: **Accuracy, sensitivity and thoroughness of STREME using default vs. optimal values of NEVAL and NREF.** The curves labeled "STREME" shows the results using the default values of NEVAL and NREF (25 and 4, respectively). The curves in Panel **a** show the percentage of ChIP-seq datasets (Y) where the best motif found by the named algorithm has motif similarity score $\geq X$, averaged over 40 ChIP-seq datasets. Each point in Panel **b** shows the percentage of times (Y) that the best motif found by the named algorithm has motif similarity score at least 5 (Tomtom p-value $\leq 10^{-5}$) when run on a primary dataset that has been diluted to a given purity (X), averaged over 40 ChIP-seq datasets. The curves in Panel **c** show the percentage of 21 reference motifs (Y) for which the named algorithm finds a motif matching it with given motif similarity score (X) or better, averaged over 20 combined datasets.



Accuracy

Sensitivity

Thoroughness

a

b

c

d

e

f

Figure S6: **Performance of motif discovery algorithms on ENCODE Gm12878 TF ChIP-seq datasets.** The top panels compare the performance of STREME with other motif discovery algorithms. The bottom panels compare using STREME with its defaults vs. using the NEVAL and NREF set to 25 and 4, respectively. The curves in Panel **a** and Panel **d** show the percentage of ChIP-seq datasets (Y) where the best motif found by the named algorithm has motif similarity score $\geq X$, averaged over 33 ChIP-seq datasets. Each point in the curves in Panel **b** and Panel **e** shows the percentage of times (Y) that the best motif found by the named algorithm has motif similarity score at least 5 (Tomtom p-value $\leq 10^{-5}$) when run on a primary dataset that has been diluted to a given purity (X), averaged over 33 ChIP-seq datasets. The curves in Panel **c** and Panel **f** show the percentage of 21 reference motifs (Y) for which the named algorithm finds a motif matching it with given motif similarity score (X) or better, averaged over 20 combined datasets.

sensitivity and thoroughness are unchanged. These results suggest that STREME's current default parameters, which give substantially faster running times than the K562-optimal parameters, are satisfactory.

2.4 Finding motifs using custom alphabets—a "novel" application

To illustrate the versatility of STREME due to its ability to use user-defined alphabets, we used it to discover the most distinguishing words in Moby Dick, a novel by Herman Melville. In many ways, this task is similar to discovering biologically important sequence motifs in biological sequences. To make the task more similar to biological motif discovery, and to make it more difficult, we remove all capitalization, spaces and punctuation from the text of the novel, and we randomly "mutate" 20% of the letters in the text to a (randomly-chosen) different letter from the English alphabet. We then create a FASTA file from the processed text, where each sentence in the novel becomes a FASTA sequence. The result is a FASTA file with 18,371 sequences ranging in length from 5 to 64 (mean length 51), containing 936,096 sequence characters in total. We also create an alphabet definition file in the simple format required by STREME, which consists of the line "`ALPHABET English`" followed by each of the 26 English letters on its own line. Although not needed here, the alphabet definition file can also specify a name and a color for each letter, as well as pairs of letters that are considered complements (as for DNA bases).

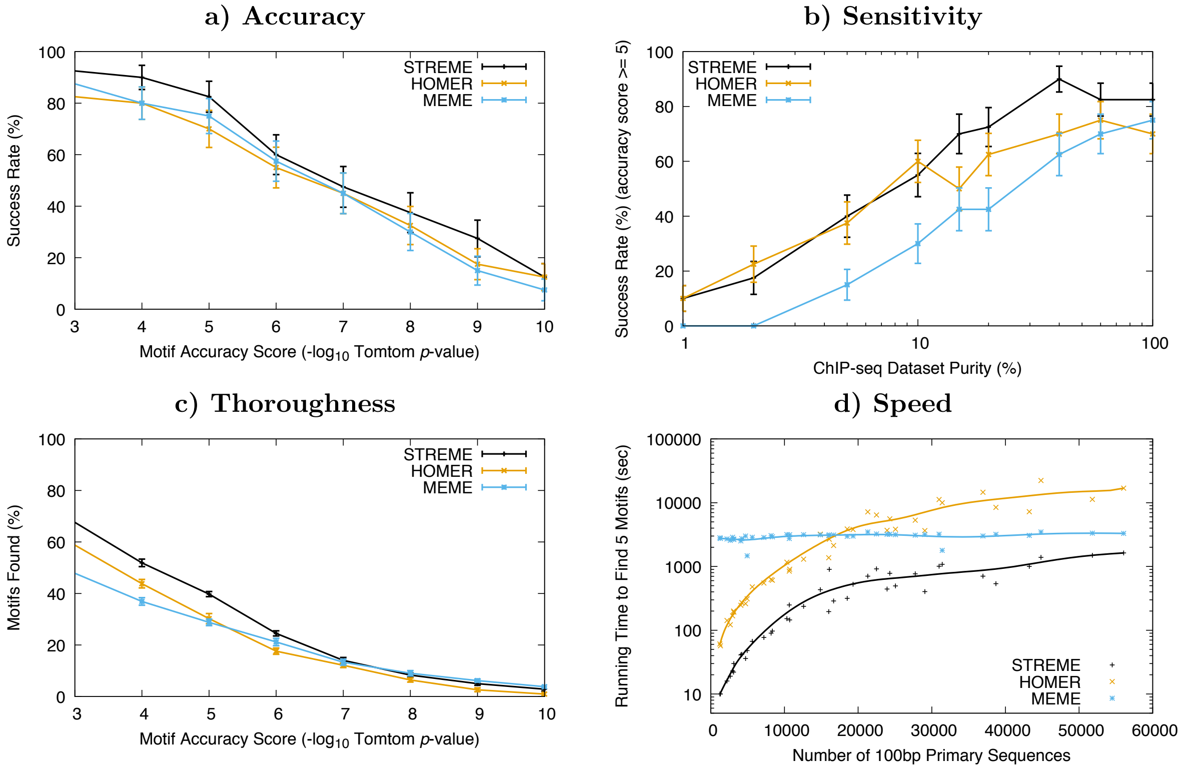
When we run STREME on the resulting FASTA file to search for motifs of widths from 3 to 12 using shuffled control sequences preserving words of width 2 (2-mers), it discovers 114 motifs with p-values less than 0.05. STREME always continues searching until the last three motifs found do not satisfy the user-specified significance level, and Table S8 shows the names, p-values and sequence logos of all 117 motifs that STREME finds in this dataset.

The most significant motif STREME finds has the consensus sequence `THEWHALE`, which is certainly an extremely distinguishing phrase in this novel about whaling during the 19th century. The vast majority of the other motifs STREME discovers are either English words (e.g., `CAPTAIN`, `AHAB`, `LEVIATHAN`), English word suffixes (`ING`, `IOUS`, `ENTLY`), or short phrases (e.g., `OFTHE`, `INTHE`, `AMOUNGTHE`). For the most part, the motifs found by STREME begin and end precisely at correct word boundaries. Occasionally the first or last letter in the word or phrase is missing from the motif. An examination of these cases reveals that this is not a bug, but an artifact of the erasing of a prior motif site that slightly overlaps a site of the truncated motif (data not shown). STREME is thus effective at discovering informative motifs that capture interesting, distinguishing words and phrases in English text.

2.5 Future Work

STREME frames the motif discovery task as optimizing a probability-based PWM to maximize classification accuracy, as measured by the Fisher exact test or the Binomial test. Ruan and Stormo [4] argues that probability models are inferior to affinity models, which can also be encoded as PWMs, because they better capture the biochemistry of transcription factor binding. Later, they used probability models of motifs (from the JASPAR TF motif database) as seeds for their DAMO algorithm, which outputs affinity-based PWMs trained on ChIP-seq peak sequences [5]. DAMO finds the optimal PWM that maximizes the AUROC (area under the receiver operating characteristic curve), an alternative measure of classification accuracy. They show that the affinity-based PWMs generated by DAMO are better at ranking bound and unbound sequences than probability-based PWMs. In future work, it would therefore be interesting to compare the AUROC of motifs from STREME and other motif finders after refinement by DAMO.

3 Supplemental Figures and Tables



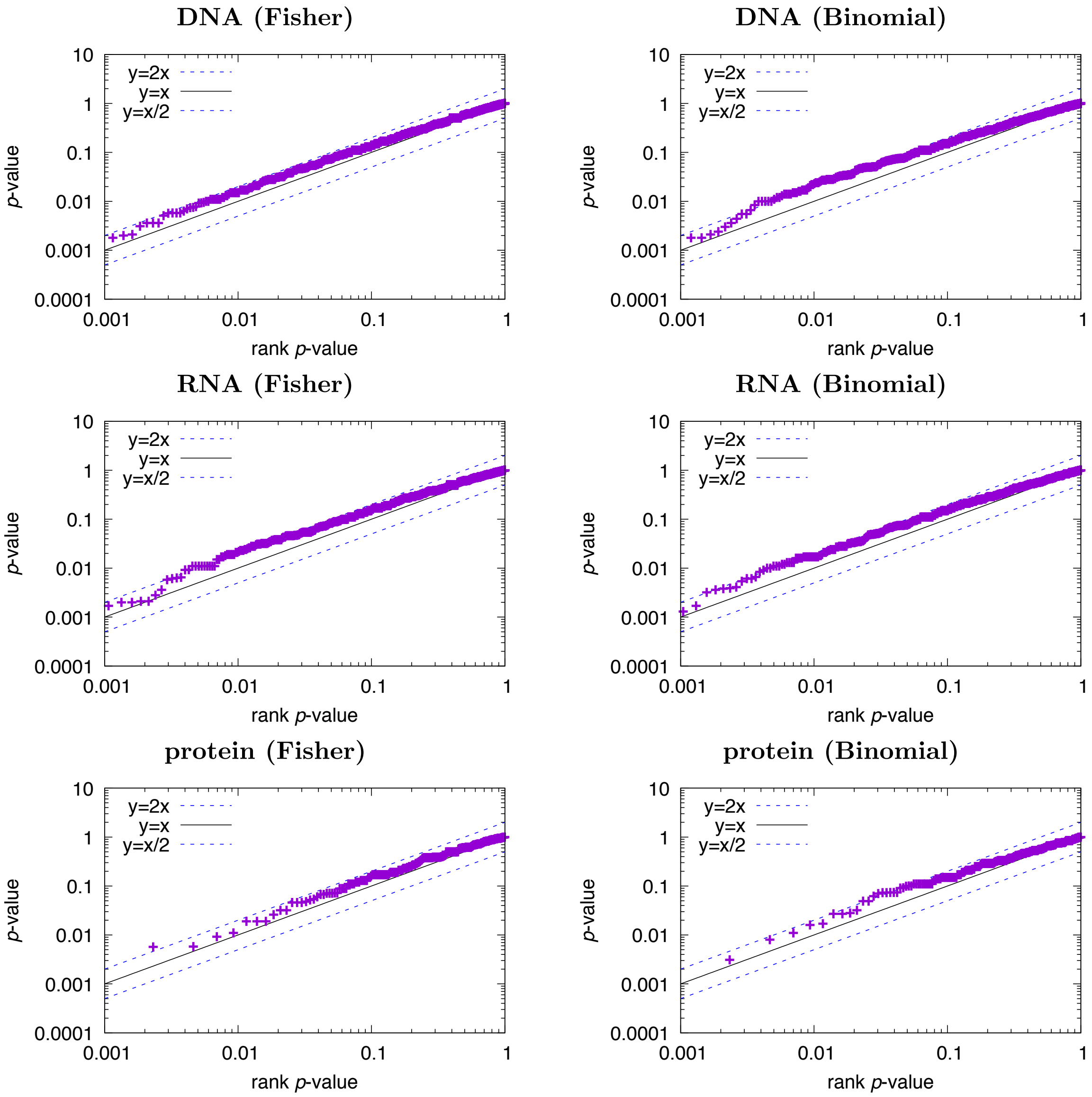
a) Accuracy

b) Sensitivity

c) Thoroughness

d) Speed

Figure S7: **Performance discovering wide motifs ($8 \leq$ width ≤ 18) in TF ChIP-seq datasets.** The curves in Panel **a** show the percentage of times (Y) the best motif found by the named algorithm has motif similarity score X or better, averaged over 40 TF ChIP-seq datasets. Each point in Panel **b** shows the percentage of times (Y) the best motif found by the named algorithm in a primary dataset that has been diluted to a purity X has motif similarity score 5 or better (Tomtom p-value $\leq 10^{-5}$), averaged over 40 TF ChIP-seq datasets. The curves in Panel **c** show the percentage of 21 reference motifs (Y) for which the named algorithm finds a motif matching it with motif similarity score X or better, averaged over 20 combined datasets, each constructed from 100 randomly-selected sequences from each of 21 TF ChIP-seq datasets. Each point in Panel **d** represents the running time (Y) of the named algorithm on one of 40 ENCODE TF ChIP-seq datasets containing X sequences. Running times are on a 4.0 GHz Intel Core i7 processor with 16GB of memory, and, for ease of interpretation, the points for each algorithm have been fit with a smooth Bezier curve.


DNA (Fisher)

DNA (Binomial)

RNA (Fisher)

RNA (Binomial)

protein (Fisher)

protein (Binomial)

Figure S8: **Q-Q accuracy plots of the p-values reported by STREME for motifs with strong support (at least 10 total sites in the hold-out set).** Each panel shows the Q-Q plot for the p-values reported by STREME when run on 10,000 datasets containing 10,000 random sequences over the alphabet given above the panel. Only motifs with at least 10 total predicted sites (primary and control sequences) in the hold-out set are included in the Q-Q plot. Primary sequences are 100 characters long. Control sequences are 100 long in the first column and 80 long in the second column of panels, causing STREME to use Fisher's exact test (first column) or the Binomial test (second column). Ideally, the points should lie along the line $y = x$. Points above that line represent conservative p-values reported by STREME. In this experiment, STREME's p-values generally lie within the lines $y = 2x$ and $y = x/2$, indicating that they are within a factor of 2 of ideal.

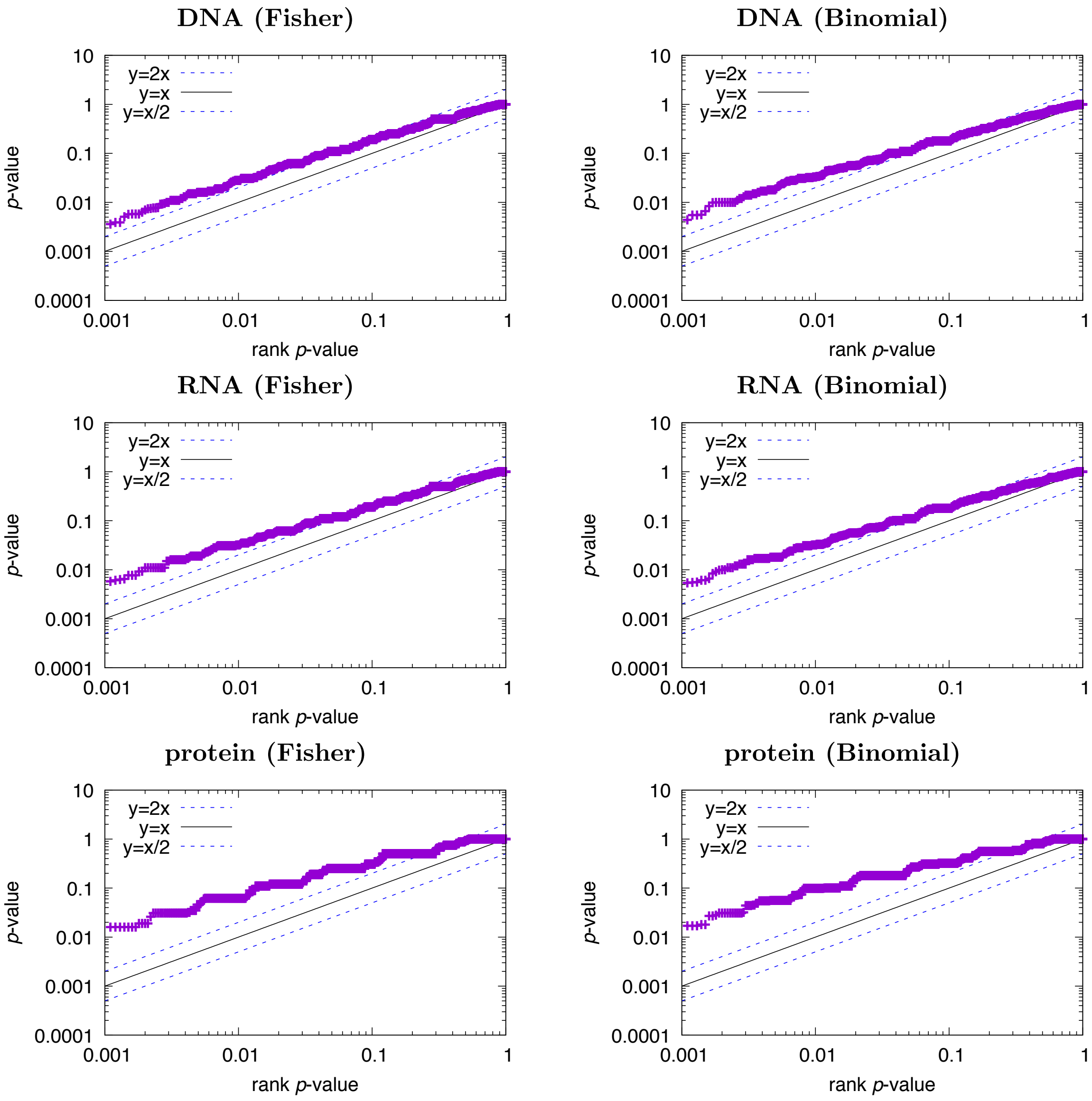
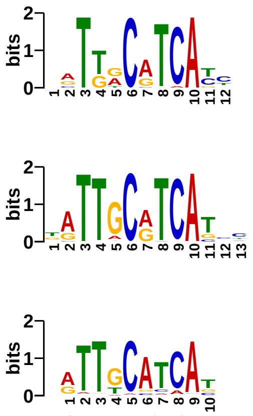
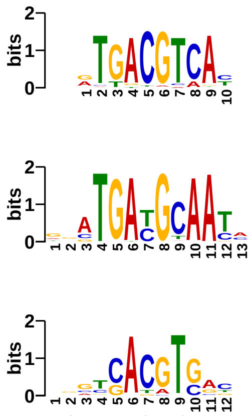
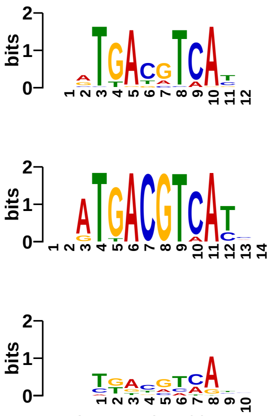
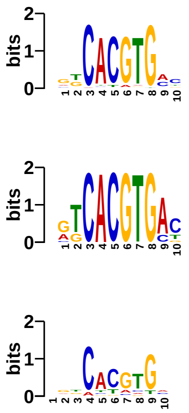
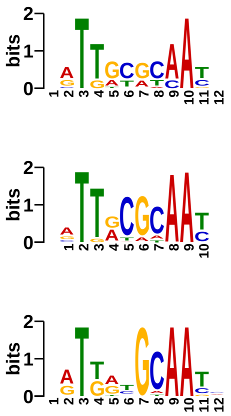
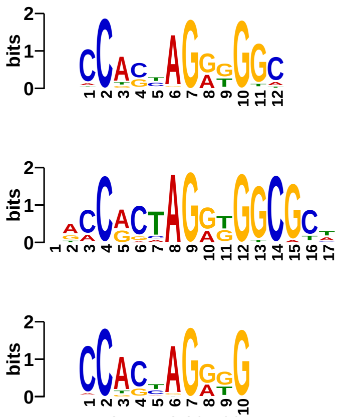


Figure S9: **Q-Q accuracy plots of the p-values reported by STREME.** Each panel shows the Q-Q plot for the p-values reported by STREME when run on 10,000 datasets containing 10,000 random sequences over the alphabet given above the panel. Control sequences are 100 long in the first column and 80 long in the second column of panels, causing STREME to use Fisher's exact test (first column) or the Binomial test (second column). Ideally, the points should lie along the line $y = x$. Points above that line represent conservative p-values reported by STREME.

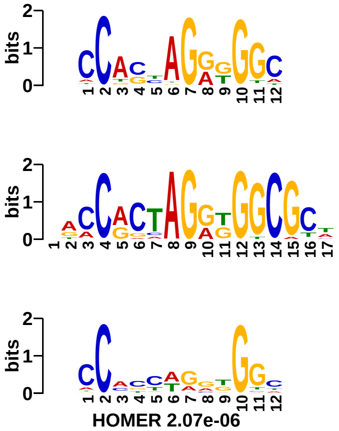
Table S6: **Comparison of STREME and HOMER ChIP-seq motifs.** The top two lines of each panel show the name of the ENCODE TF ChIP-seq dataset and the name of the SELEX motif (from Jolma *et al.* [1]) that we used for evaluating that experiment. The three sequence logos show the STREME motif (top) and the HOMER motif (bottom) aligned to the reference motif (center). Above and below the aligned logos are the names of the two algorithms and the accuracy (Tomtom *p*-value) of the motif discovered by the named algorithm. (A smaller Tomtom *p*-value indicates that the discovered motif is more similar to the reference motif.) The motif found by STREME is more accurate than that found by HOMER for 29 out of the 40 TF ENCODE ChIP-seq datasets (done in K562 cells) for which there is a high-throughput SELEX reference motif.

HaibK562Atf3V0416101 ATF4_DBD **STREME 4.81e-10**  **HOMER 1.40e-07**	SydhK562Atf3 ATF4_DBD **STREME 6.22e-03**  **HOMER 1.25e-01**	SydhK562Atf106325 ATF7_DBD **STREME 3.48e-10**  **HOMER 4.85e-03**
SydhK562Bhlhe40nb100Iggrab BHLHE41_full **STREME 1.67e-09**  **HOMER 1.11e-04**	SydhK562CebpbIggrab CEBPB_full **STREME 3.47e-10**  **HOMER 1.09e-07**	BroadK562Ctcf CTCF_full **STREME 1.56e-10**  **HOMER 8.86e-09**

UtaK562Ctcf

CTCF_full

STREME 1.31e-10

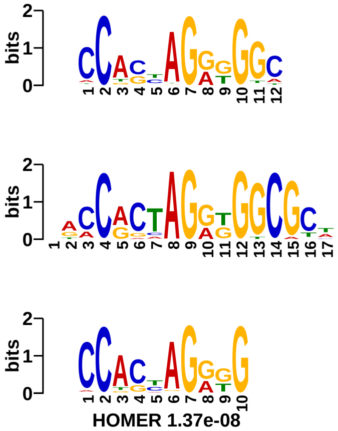


HOMER 2.07e-06

UwK562Ctcf

CTCF_full

STREME 1.92e-10

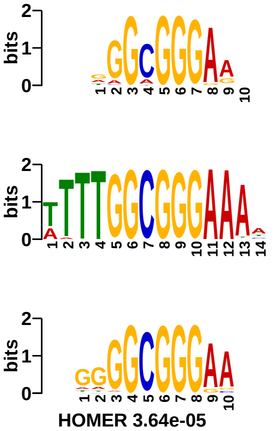


HOMER 1.37e-08

HaibK562E2f6V0416102

E2F7_DBD

STREME 1.12e-05

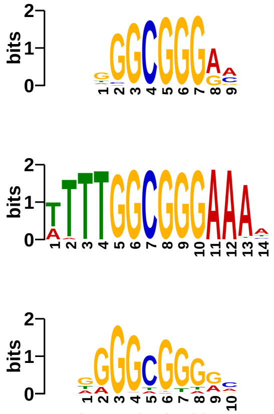


HOMER 3.64e-05

SydhK562E2f4Ucd

E2F7_DBD

STREME 2.23e-05

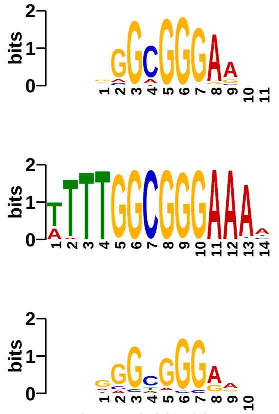


HOMER 9.76e-03

SydhK562E2f6Ucd

E2F7_DBD

STREME 7.98e-06

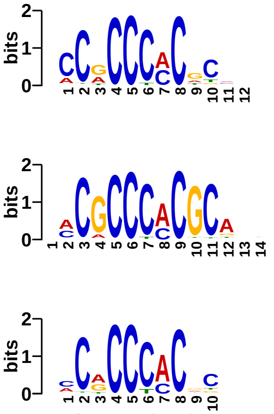


HOMER 2.32e-03

HaibK562Egr1V0416101

EGR1_DBD

STREME 2.11e-09

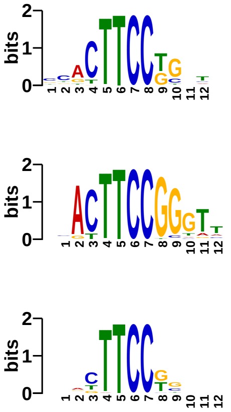


HOMER 1.94e-07

HaibK562Elf1sc631V0416102

ELF1_DBD

STREME 1.12e-07

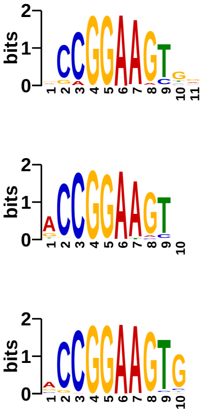


HOMER 1.27e-06

SydhK562Elk112771Iggrab

ELK1_DBD_2

STREME 2.82e-09

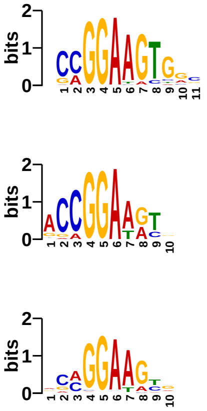


HOMER 1.89e-08

HaibK562Ets1V0416101

ETS1_DBD_1

STREME 4.13e-06

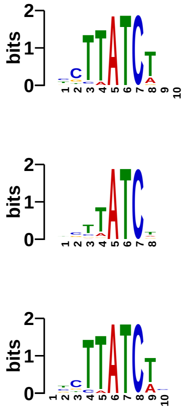


HOMER 6.31e-08

HaibK562Gata2sc267Pcr1x

GATA3_DBD

STREME 1.81e-06

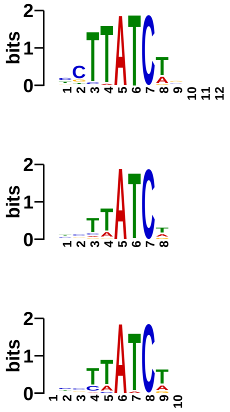


HOMER 1.17e-06

SydhK562Gata1Ucd

GATA4_DBD

STREME 1.88e-06

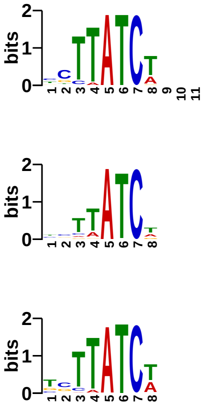


HOMER 7.99e-12

SydhK562Gata2Ucd

GATA4_DBD

STREME 2.55e-07

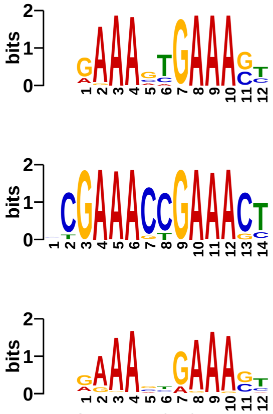


HOMER 3.66e-08

SydhK562Irf1Ifna6h

IRF8_full

STREME 1.61e-06

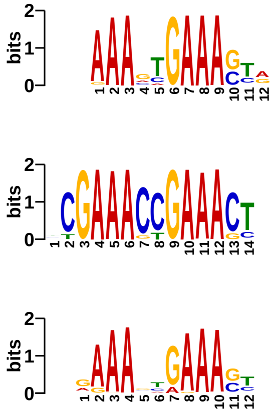


HOMER 1.10e-05

SydhK562Irf1Ifng30

IRF8_full

STREME 6.39e-06

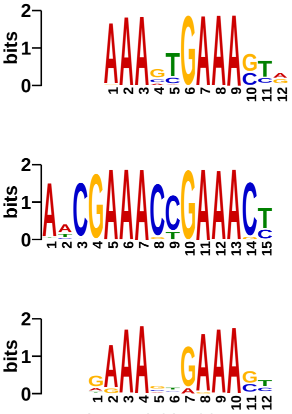


HOMER 7.52e-06

SydhK562Irf1Ifna30

IRF9_full

STREME 9.68e-06

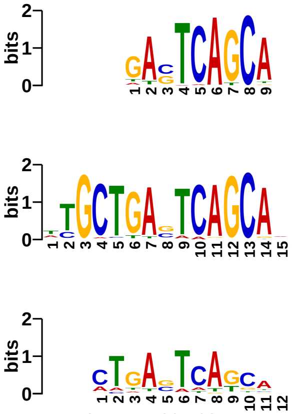


HOMER 9.08e-06

SydhK562MaffIggrab

MAFF_DBD

STREME 2.74e-09

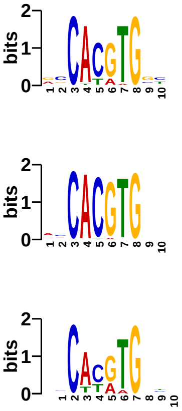


HOMER 1.30e-09

HaibK562MaxV0416102

MAX_DBD_2

STREME 2.91e-10

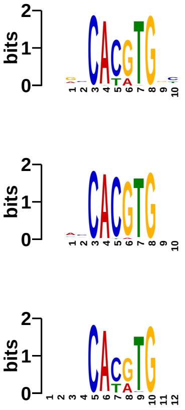


HOMER 7.07e-08

SydhK562MaxIggrab

MAX_DBD_2

STREME 2.18e-09

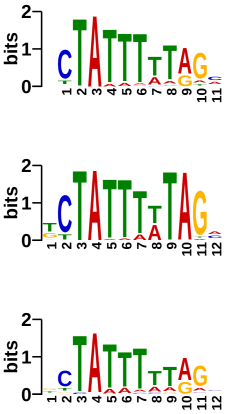


HOMER 3.81e-09

HaibK562Mef2aV0416101

MEF2A_DBD

STREME 9.59e-11

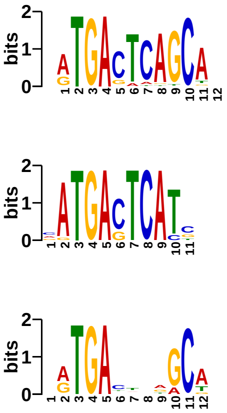


HOMER 1.10e-09

SydhK562Nfe2

NFE2_DBD

STREME 6.43e-05

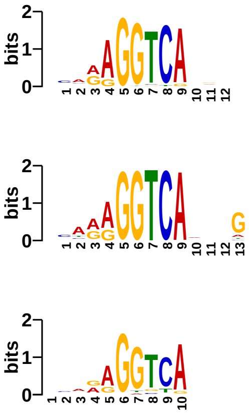


HOMER 6.80e-03

HaibK562Nr2f2sc271940V0422111

NR2F1_DBD_3

STREME 1.30e-14

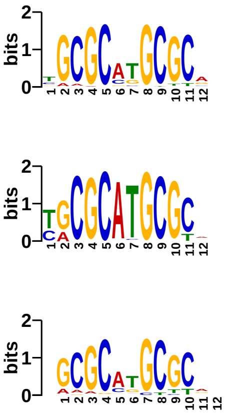


HOMER 7.34e-08

SydhK562Nrf1Iggrab

NRF1_full

STREME 1.36e-10

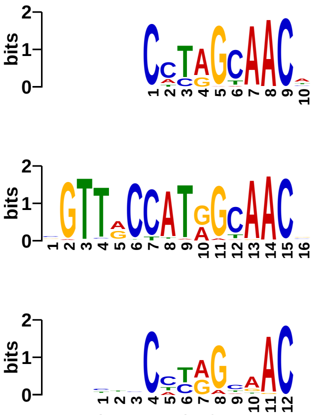


HOMER 1.93e-09

SydhK562Rfx5Iggrab

RFX5_DBD_2

STREME 1.36e-05

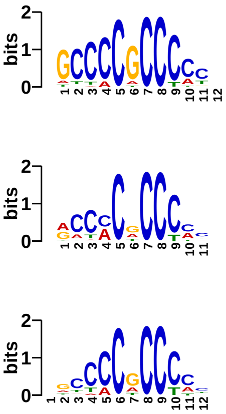


HOMER 1.58e-04

HaibK562Sp1Pcr1x

SP1_DBD

STREME 3.29e-08

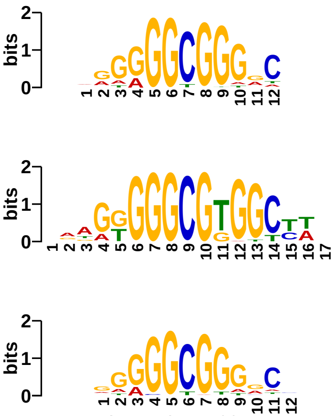


HOMER 1.92e-13

HaibK562Sp2sc643V0416102

SP4_full

STREME 1.29e-06

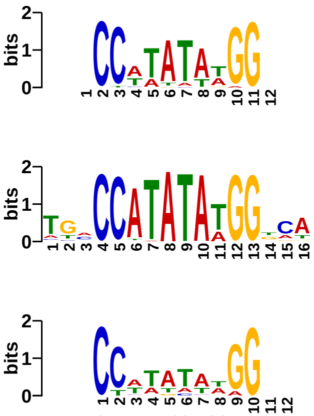


HOMER 2.14e-06

HaibK562SrfV0416101

SRF_full

STREME 2.09e-08

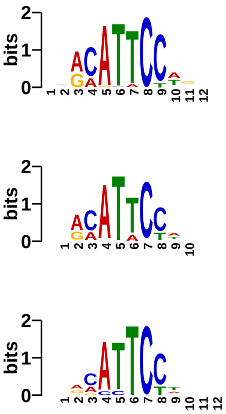


HOMER 5.08e-06

HaibK562Tead4sc101184V0422111

TEAD4_DBD

STREME 7.67e-10

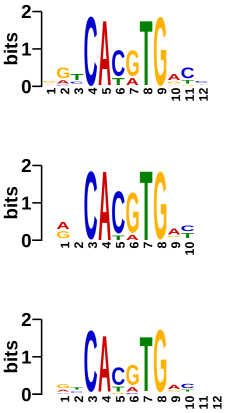


HOMER 1.45e-09

HaibK562Usf1V0416101

USF1_DBD

STREME 1.60e-11

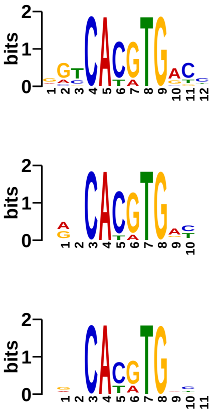


HOMER 5.23e-11

SydhK562Usf2Iggrab

USF1_DBD

STREME 1.86e-10

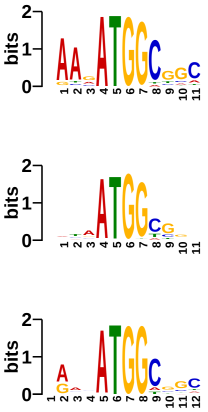


HOMER 4.02e-11

HaibK562Yy1V0416101

YY1_full

STREME 2.98e-06

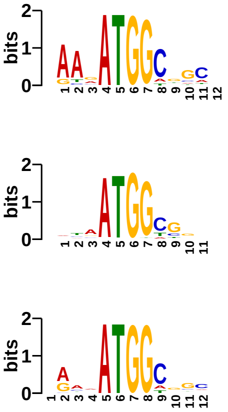


HOMER 2.40e-09

HaibK562Yy1V0416102

YY1_full

STREME 6.86e-07

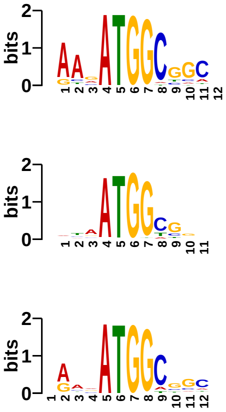


HOMER 1.82e-10

SydhK562Yy1Ucd

YY1_full

STREME 2.31e-06

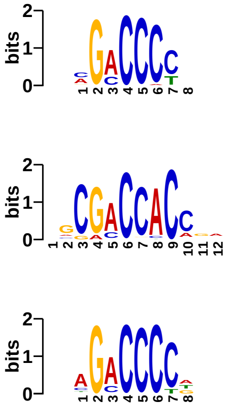


HOMER 2.52e-09

HaibK562Zbtb7asc34508V0416101

ZBTB7A_DBD

STREME 9.52e-03

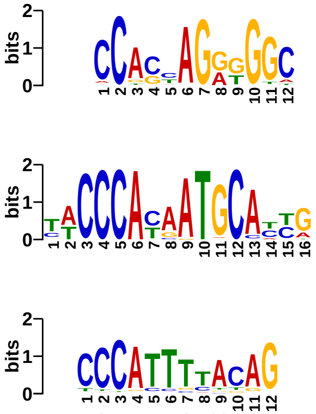


HOMER 1.82e-02

SydhK562Znf143Iggrab

ZNF143_DBD

STREME 2.09e-01



HOMER 7.78e-02

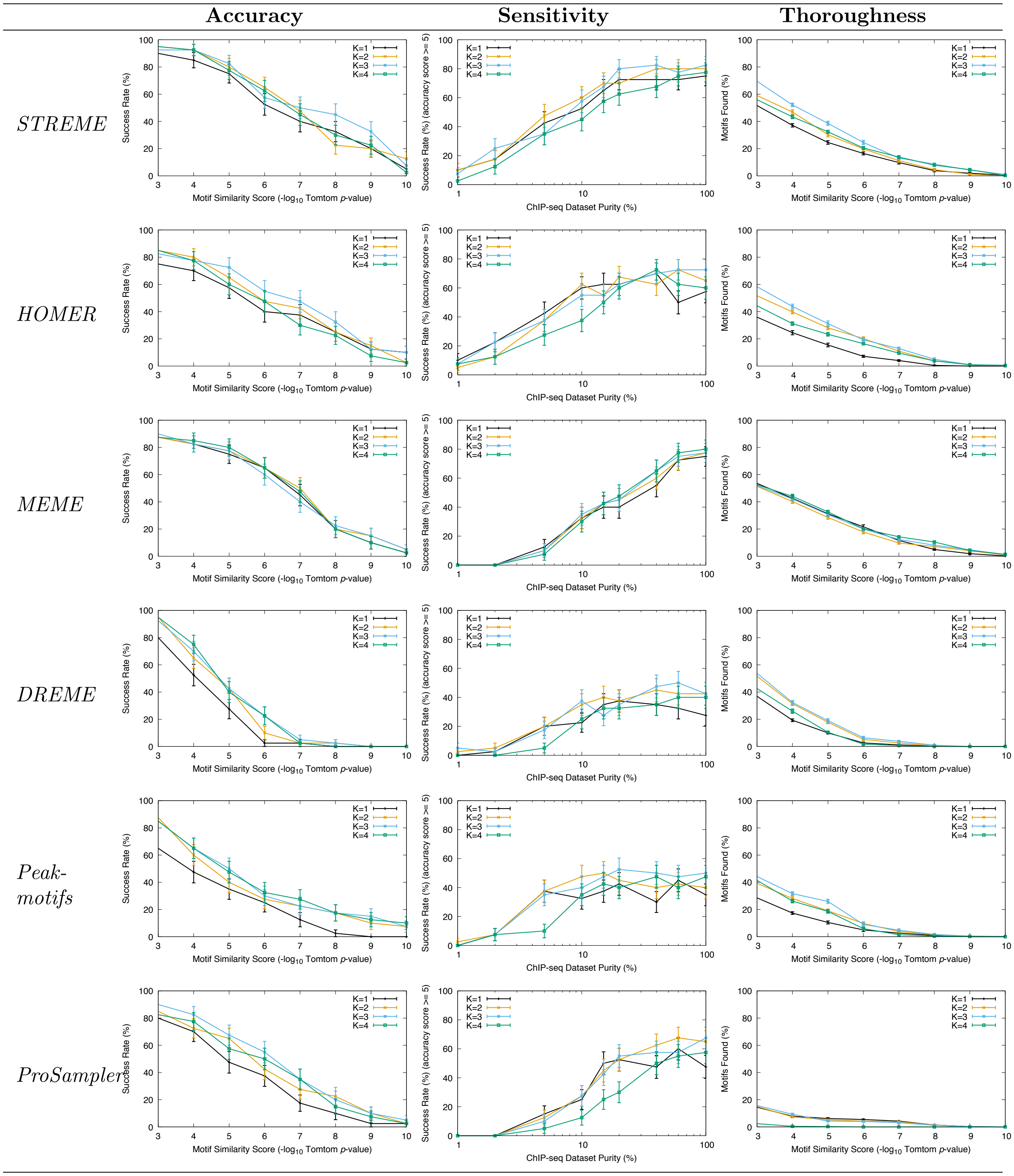
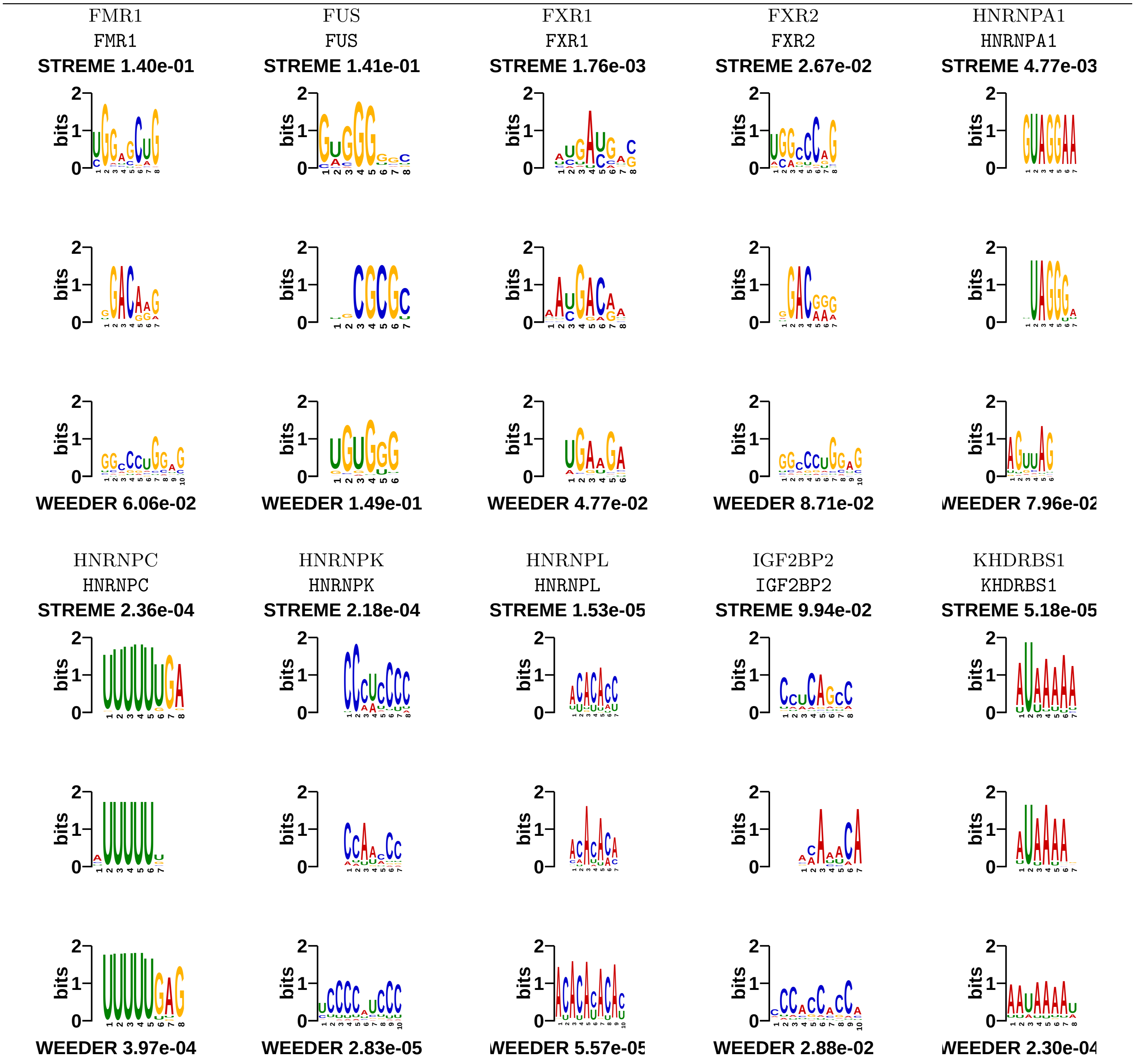


Figure S10: **Choice of shuffling k-mer size (k) on motif discovery in ChIP-seq datasets.** The panels show the accuracy, sensitivity and thoroughness of different motif discovery algorithms as a function of the size of words (k) whose frequencies are preserved when shuffling the primary sequences to create the control sequences. STREME, MEME and Peak-motifs also use $k - 1$ as the order of the Markov model they create internally from the input sequences. (Weeder does not use a control set.) The contents of the plots in the three columns are described in the captions to Figures 1, 2 and 3, respectively. The sensitivity plots use a motif similarity score threshold of 5 (Tomtom p-value $\leq 10^{-5}$).

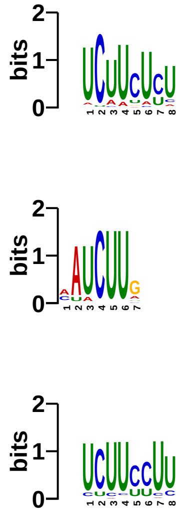
Table S7: **Comparison of STREME and Weeder eCLIP motifs.** The top two lines of each panel show the name of the ENCODE RNA-binding protein eCLIP dataset and the name of the RNAcompete motif (from Ray *et al.* [3]) that we used for evaluating that experiment. The three sequence logos show the STREME motif (top) and the Weeder motif (bottom) aligned to the reference motif (center). Above and below the aligned logos are the names of the two algorithms and the accuracy (Tomtom p-value) of the motif discovered by the named algorithm. (A smaller Tomtom p-value indicates that the discovered motif is more similar to the reference motif.)



MATR3

MATR3

STREME 8.45e-03

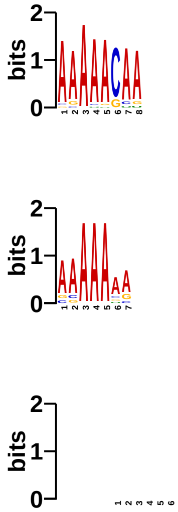


WEEDER 1.30e-02

PABPC4

PABPC4

STREME 5.77e-04

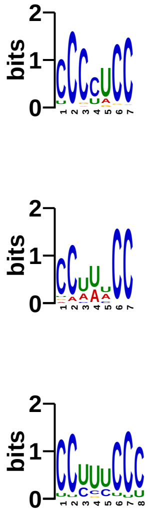


WEEDER 9.81e-01

PCBP1

PCBP1

STREME 1.67e-04

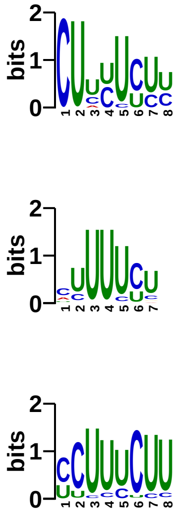


WEEDER 2.76e-06

PTBP1

PTBP1

STREME 7.50e-05

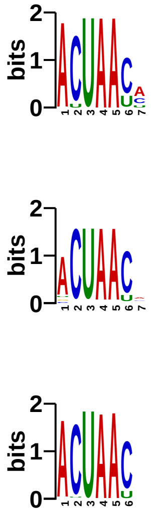


WEEDER 2.69e-04

QKI

QKI

STREME 2.34e-07

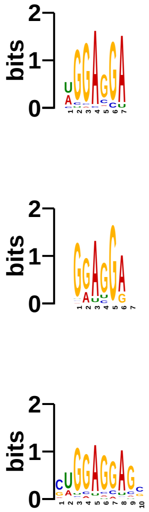


WEEDER 2.99e-06

SRSF1

SRSF1

STREME 1.83e-04

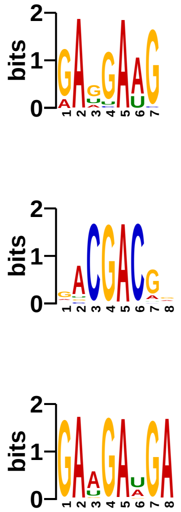


WEEDER 1.22e-04

SRSF7

SRSF7

STREME 5.09e-02

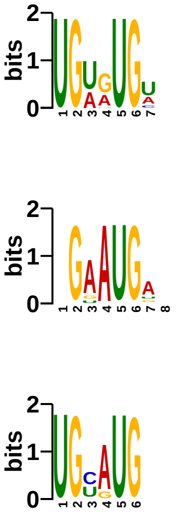


WEEDER 3.70e-02

TARDBP

TARDBP

STREME 7.79e-03

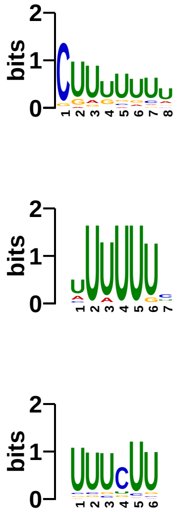


WEEDER 3.63e-03

TIA1

TIA1

STREME 1.15e-03

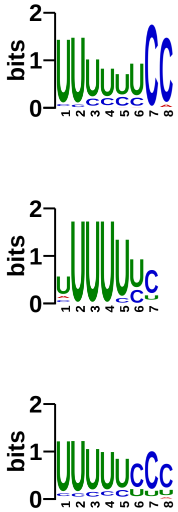


WEEDER 3.16e-03

U2AF2

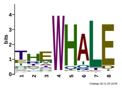
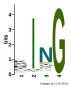
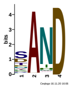
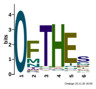
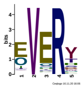
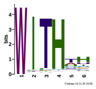
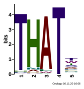
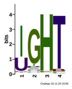
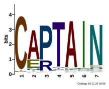
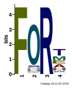
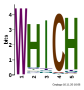
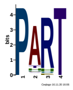
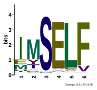
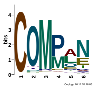
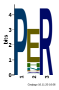
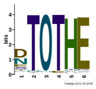
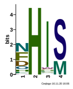
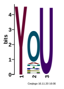
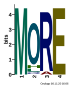
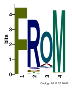
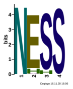
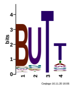
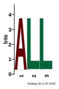
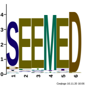
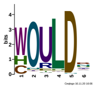
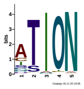
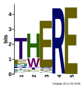
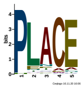
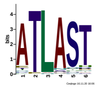
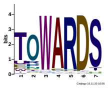
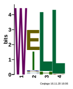
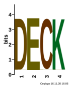
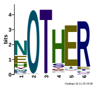
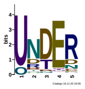
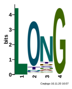
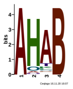
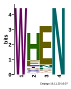
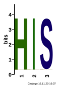
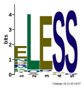
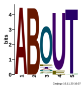
U2AF2

STREME 4.65e-05

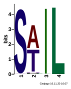


WEEDER 2.17e-05

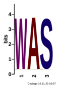
Table S8: **Motifs found by STREME in the corrupted text of the novel Moby Dick.** The table shows names, p-values and sequence logos of the 117 motifs discovered by STREME in the text of Moby Dick, from which we removed all spacing, punctuation and capitalization, and to which we added 20% noise. We run STREME with the minimum motif width set to 3, the maximum width set to 12, and the size of k-mers to preserve set to 2.

1-THEWHALE 6.8e-017 	2-KING 6.5e-016 	3-SAND 6.8e-016 	4-OFTHES 9.4e-013 	5-EVERY 1.5e-011 
6-WITHTH 1.9e-010 	7-THATT 3.8e-010 	8-IGHT 2.6e-009 	9-CAPTAIN 8.8e-009 	10-FORT 7.4e-008 
11-WHICH 7.8e-008 	12-PART 2.4e-007 	13-IMSELF 2.7e-007 	14-COMPAN 2.1e-006 	15-PER 2.9e-006 
16-DTOTHE 3.4e-006 	17-NHIS 5.5e-006 	18-YOU 7.8e-006 	19-MORE 1.1e-005 	20-FROM 1.4e-005 
21-NESS 1.9e-005 	22-BUTT 2.5e-005 	23-ALL 4.3e-005 	24-SEEMED 5.0e-005 	25-WOULDB 6.8e-005 
26-ATION 7.5e-005 	27-THERE 1.2e-004 	28-PLACE 1.3e-004 	29-ATLAST 1.3e-004 	30-TOWARDS 1.3e-004 
31-WELL 1.4e-004 	32-DECK 1.5e-004 	33-NOTHER 1.8e-004 	34-UNDER 2.0e-004 	35-LONG 2.1e-004 
36-AHAB 3.5e-004 	37-WHEN 3.9e-004 	38-HIS 4.0e-004 	39-ELESS 4.0e-004 	40-ABOUT 4.0e-004 

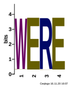
41-SAIL
4.2e-004



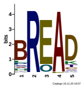
42-WAS
4.9e-004



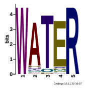
43-WERE
5.0e-004



44-BREAD
5.1e-004

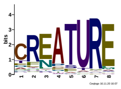


45-WATER
5.2e-004

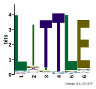


46-INSTANCE
5.2e-004

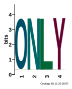
47-CREATURE
5.2e-004



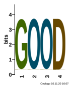
48-LITTLE
5.3e-004



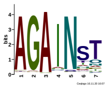
49-ONLY
7.1e-004



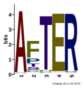
50-GOOD
7.1e-004



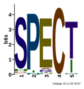
51-AGAINST
7.2e-004



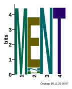
52-AFTER
7.4e-004



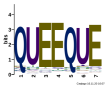
53-SPECT
8.2e-004



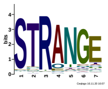
54-MENT
8.4e-004



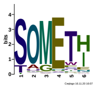
55-QUEEQUE
9.9e-004



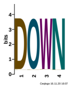
56-STRANGE
9.9e-004



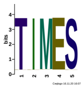
57-SOMETH
1.3e-003



58-DOWN
1.4e-003

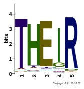


59-TIMES
2.1e-003

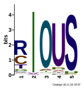


60-HARPOONEER
2.1e-003

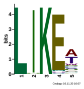
61-THEIR
2.2e-003



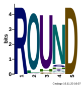
62-RIOUS
2.2e-003



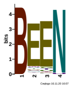
63-LIKEA
2.4e-003



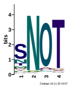
64-ROUND
2.4e-003



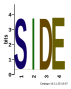
65-BEEN
2.9e-003



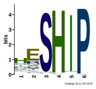
66-SNOT
3.0e-003



67-SIDE
3.9e-003

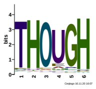


68-HESHIP
3.9e-003

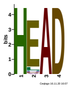


69-LEVIATHAN
4.1e-003

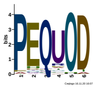
70-THOUGH
4.5e-003



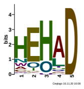
71-HEAD
5.3e-003



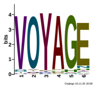
72-PEQUOD
6.2e-003



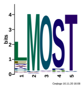
73-HEHAD
6.4e-003



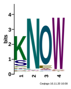
74-VOYAGE
8.2e-003



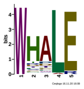
75-LMOST
8.8e-003



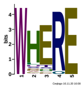
76-KNOW
9.0e-003



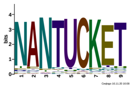
77-WHALE
1.1e-002



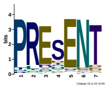
78-WHERE
1.1e-002



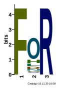
79-NANTUCKET
1.1e-002



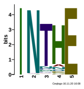
80-PRESENT
1.2e-002



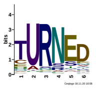
81-FOR
1.4e-002



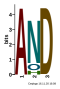
82-INTHE
1.4e-002



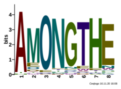
83-TURNED
1.4e-002



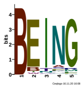
84-AND
1.4e-002



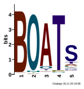
85-AMONGTHE
1.6e-002



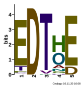
86-BEING
1.9e-002



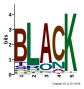
87-BOATS
1.9e-002



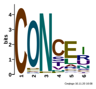
88-EDTHE
1.9e-002



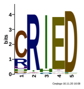
89-BLACK
2.0e-002



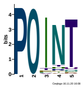
90-CONCEI
2.0e-002



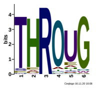
91-CRIED
2.0e-002



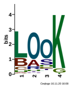
92-POINT
2.0e-002



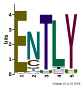
93-THROUG
2.0e-002



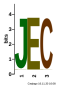
94-LOOK
2.1e-002



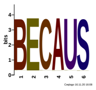
95-ENTLY
3.0e-002



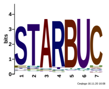
96-JEC
3.2e-002



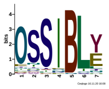
97-BECAUS
3.2e-002



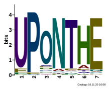
98-STARBUC
3.2e-002



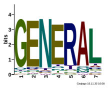
99-OSSIBLY
3.2e-002



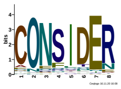
100-UPONTHE
3.2e-002



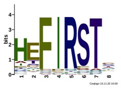
101-GENERAL
3.2e-002



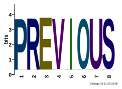
102-CONSIDER
3.2e-002



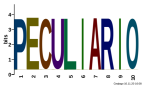
103-HEFIRSTC
3.2e-002



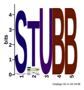
104-PREVIOUS
3.2e-002



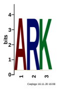
105-PECULIARIO
3.2e-002



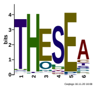
106-STUBB
3.4e-002



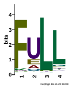
107-ARK
3.5e-002



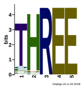
108-THESEA
3.5e-002



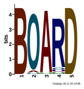
109-FULL
3.6e-002



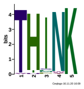
110-THREE
3.7e-002



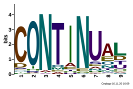
111-BOARD
3.7e-002



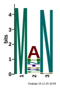
112-THINK
3.7e-002



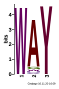
113-CONTINUAL
3.7e-002



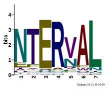
114-MAN
4.6e-002



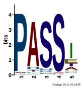
115-WAY
5.4e-002



116-NTERVAL
1.5e-001



117-PASSI
1.7e-001



References

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

[2] T. Lesluyes, J. Johnson, P. Machanick, and T. L. Bailey. Differential motif enrichment analysis of paired ChIP-seq experiments. *BMC Genomics*, 15:752, 2014.

[3] D. Ray, H. Kazan, K. B. Cook, M. T. Weirauch, H. S. Najafabadi, X. Li, S. Gueroussov, M. Albu, H. Zheng, A. Yang, H. Na, M. Irimia, L. H. Matzat, R. K. Dale, S. A. Smith, C. A. Yarosh, S. M. Kelly, B. Nabet, D. Mecenas, W. Li, R. S. Laishram, M. Qiao, H. D. Lipshitz, F. Piano, A. H. Corbett, R. P. Carstens, B. J. Frey, R. A. Anderson, K. W. Lynch, L. O. F. Penalva, E. P. Lei, A. G. Fraser, B. J. Blencowe, Q. D. Morris, and T. R. Hughes. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–177, Jul 2013.

[4] S. Ruan and G. D. Stormo. Inherent limitations of probabilistic models for protein-dna binding specificity. *PLoS computational biology*, 13:e1005638, July 2017.

[5] S. Ruan and G. D. Stormo. Comparison of discriminative motif optimization using matrix and dna shape-based models. *BMC bioinformatics*, 19:86, Mar. 2018.

[6] E. L. Van Nostrand, G. A. Pratt, A. A. Shishkin, C. Gelboin-Burkhart, M. Y. Fang, B. Sundararaman, S. M. Blue, T. B. Nguyen, C. Surka, K. Elkins, R. Stanton, F. Rigo, M. Guttman, and G. W. Yeo. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature methods*, 13:508–514, June 2016.