SUPPLEMENTARY NOTES

Supplementary Note 1. The MoSBAT algorithm and possible extensions

For comparing two motifs, MoSBAT starts by converting the motifs to position-specific affinity matrices (PSAMs) (Foat, et al., 2006). For each motif position in the PSAM, the most preferred nucleotide obtains a value of 1.0, and each of the other nucleotides obtains a value equal to its frequency of occurrence divided by the frequency of occurrence of the most preferred nucleotide. In other words, a PSAM is a position-frequency matrix (PFM) that is normalized so that the maximum value at each motif position is 1.0.

Next, each PSAM is used to scan a set of N random sequences, resulting in two vectors of length N containing the PSAM scores for each sequence. The PSAM score for each sequence is calculated by taking the sum of binding probabilities across the sequence, with binding probabilities at equilibrium calculated as described before (Zhao and Stormo, 2011):

$$S = \sum_{i=1}^{L-w+1} \frac{1}{1 + \dfrac{1}{\prod_{j=1}^{w} P_j\left(s_{i+j-1}\right)}} \qquad (1)$$

Here, L is the length of the sequence, w is the width of the PSAM, s_x is the nucleotide at position x of the sequence, and $P_j(s_x)$ is the value associated with that nucleotide at position j of the PSAM.

Once the PSAM scores for each of the two motifs for each of the N sequences are calculated, the Pearson correlation of the resulting two vectors is taken as the similarity of the binding "affinity" profile of the two motifs (*i.e.* MoSBAT-a score). Alternatively, the PSAM scores can first be converted to binding "energy" scores by taking the logarithm of affinities, then the Pearson correlations can be calculated, resulting in the MoSBAT-e score.

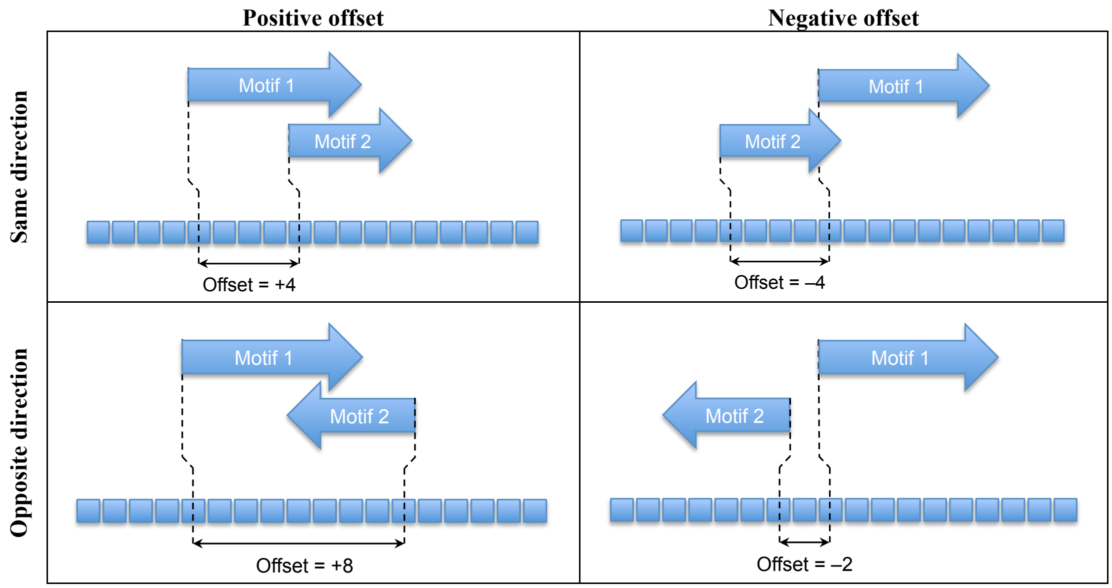
To accurately estimate the Pearson correlations, the number of random sequences, N, should be large (recommended: >50,000) so as to minimize variance in MoSBAT scores (discussed in Supplementary Note 3). Each sequence is by default generated based on a mononucleotide uniform distribution (¼ A, ¼ C, ¼ G, ¼ T) with a user-specified length L (recommended: <100 nt; see Supplementary Note 3). In some cases, users may wish to compare their motifs using genomic sequences, sequences with different GC content, or sequences with a particular set of dinucleotide frequencies. These sequence sets can be easily incorporated into the MoSBAT program, as described in the README file that is provided along with the open-source distribution of MoSBAT (https://github.com/csglab/MoSBAT).

We note that MoSBAT does not explicitly calculate a similarity threshold, or significance *p*-value for motif similarity, as almost any Pearson correlation coefficient estimated with high-dimensional data is significant. Cutoffs could be established with a sensitivity/specificity analysis based on classification tasks such as associating the *in vivo* motifs to *in vitro* motifs. In our work, we note that the MoSBAT-a scores > 0.6 and MoSBAT-e scores > 0.8 are reasonable thresholds for similar motifs. Alternatively, MoSBAT scores can be used to calculate empirical *p*-values based on large collections of random motifs with similar characteristics to the user query set. This procedure has been previously described (Mahony and Benos, 2007; Sandelin and Wasserman, 2004), and we have outlined a possible implementation in the README file available at https://github.com/csglab/MoSBAT.

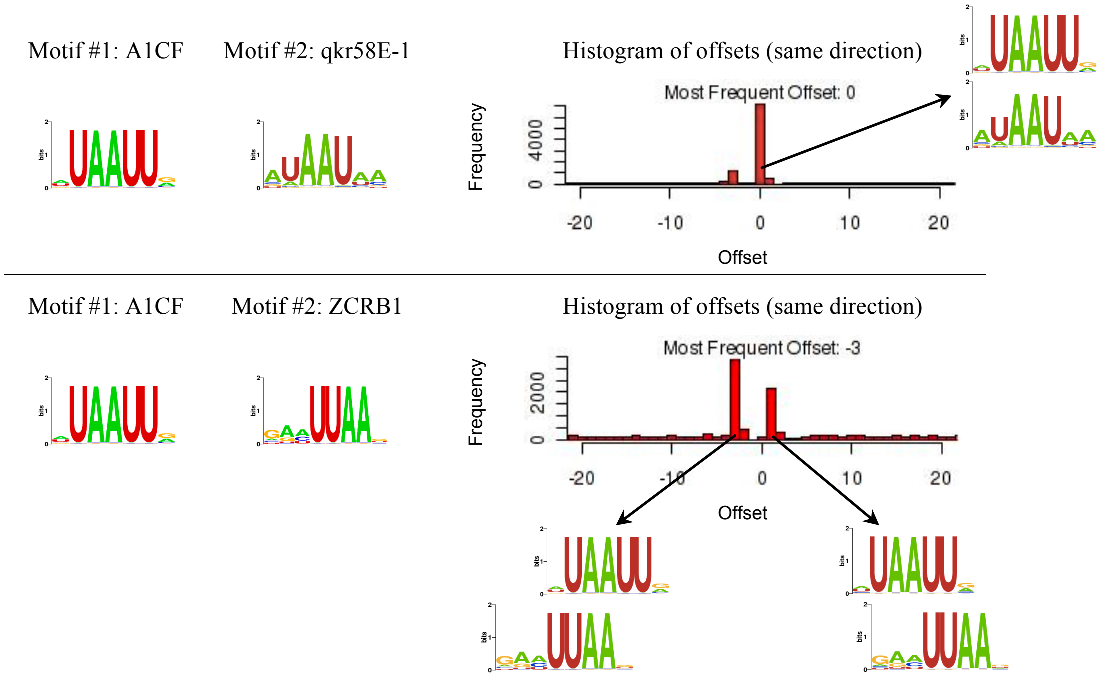
Supplementary Note 2. Calculating motif alignments

Although MoSBAT is an alignment-independent approach for measuring motif similarity, the implementation of MoSBAT is able to report a representation of the alignment of the queried motifs. However, instead of showing only one possible alignment, MoSBAT shows the distribution of possible alignments as a histogram. In other words, it shows how frequently the two motifs occur at a particular distance from each other in a given set of sequences (*e.g.* for a set of randomly generated sequences, which is used by default). This is particularly useful for cases where a shorter motif could align to multiple different positions in a longer motif, where low-complexity motifs (such as poly-U motifs) can be aligned in multiple ways, or where palindromic motifs are aligned.

To calculate the motif alignments, MoSBAT first reports the location of the best match in each sequence, which is the position that results in the maximum PSAM score. Next, offsets between two motifs are calculated for motif occurrences that are in the same direction and motif occurrences that are at opposite directions separately. The distribution of the offsets across all the tested sequences is reported as a histogram (forward and reverse histograms). In each histogram, the most frequent offset is also shown. Below are some schematic examples showcasing how the offset of two motif occurrences is calculated.



Two example offset histograms:


Motif #1: A1CF
Motif #2: qkr58E-1
Histogram of offsets (same direction)

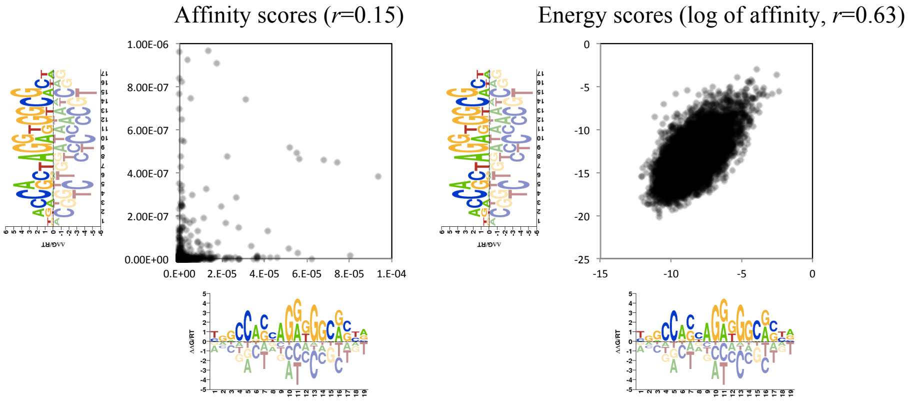
Motif #1: A1CF
Motif #2: ZCRB1
Histogram of offsets (same direction)

Supplementary Note 3. Choice of parameters for MoSBAT

Since MoSBAT uses randomly generated sequences to calculate motif similarity, we sought to explore how properties of the motifs themselves along with the two user-specified parameters *L* (sequence length) and *N* (sequence number) affect the variance in MoSBAT scores (Supplementary Figure 5). We noted that, in the range that we examined, sequence length has a smaller effect on MoSBAT score variance compared to sequence number. Based on these results we suggest that *N* be set to at least 50,000 random sequences, with larger numbers further decreasing the variance.

Furthermore, we note that the use of long sequences may result in loss of information, since the PSAM score for extremely long sequences is mainly determined by the GC content of the sequence and the motif, rather than by the specific sequence of the motif. This effect, however, is only tangible when the sequence length is extremely long relative to the information content of the motif. For example, an RNA motif that is 6nt wide (12 bit information) occurs in ~2.5% of random sequences that are 100nt long, and therefore the affinity profile of such a motif across such sequences would be informative of the underlying sequence preference. However, the same motif is expected to occur at least once in almost any sequence that is 4,000nt long, and therefore the affinity score profile across such sequences would not be informative. We recommend the use of the minimum possible sequence length for comparison of two motifs in order to maximally retain the information about the motif sequence in the PSAM score profile – this minimum possible length is approximately the sum of the lengths of the two motifs that are being compared.

In addition, there are differences in variance between MoSBAT-a and MoSBAT-e scores. Affinity scores heavily penalize sequences that do not match the consensus, and tend to cluster near zero even with a few mismatches between the motif and the sequence. Therefore, in order to obtain enough variability in the affinity profile (i.e. enough non-zero scores), a larger number of sequences should be analyzed. Indeed, the number of occurrences of perfect or near-perfect matches drives most of the variance in the affinity values (Equation 1 in Supplementary Note 1). This particularly makes the comparisons that include longer motifs, which naturally occur less frequently in random sequences, more variable when MoSBAT-a is used, since perfect or near-perfect motif matches are less probable for such motifs (Supplementary Figure 5). In contrast, the distribution of the logarithm of affinity (energy) is smoother, as the logarithmic transformation expands the variance among the near-zero values, allowing for more information to be extracted from the energy score distribution. An example is shown here, comparing the affinity score scatterplot of two different CTCF motifs in the linear scale and in the logarithmic scale:



Affinity scores (*r*=0.15)

Energy scores (log of affinity, *r*=0.63)

To accurately calculate MoSBAT-a scores with long motifs, we recommend setting *N* as high as possible. However, achieving the required number of sequences to properly calculate MoSBAT-a scores for very long motifs might not be practical. For example, for a motif with a length of 10, approximately 1,000,000 sequences of length 100 should be analyzed in order to have ~100 sequences with perfect motif matches, which may not be computationally feasible for most applications. However, we note that naturally occurring sequences, such as different regions of the genomic sequence, are not random, and may have been selected for presence of such long motifs. Therefore, the use of genomic sequences instead of random sequences could be beneficial in such cases (see Supplementary Note 1).

Overall, due to the smaller variance in MoSBAT-e scores, and its better performance in the majority of benchmarking tests, we recommend using MoSBAT-e scores for most tasks, with *N* (number of sequences) set as high as possible and *L* (sequence length) set as low as possible.

References for Supplementary Notes

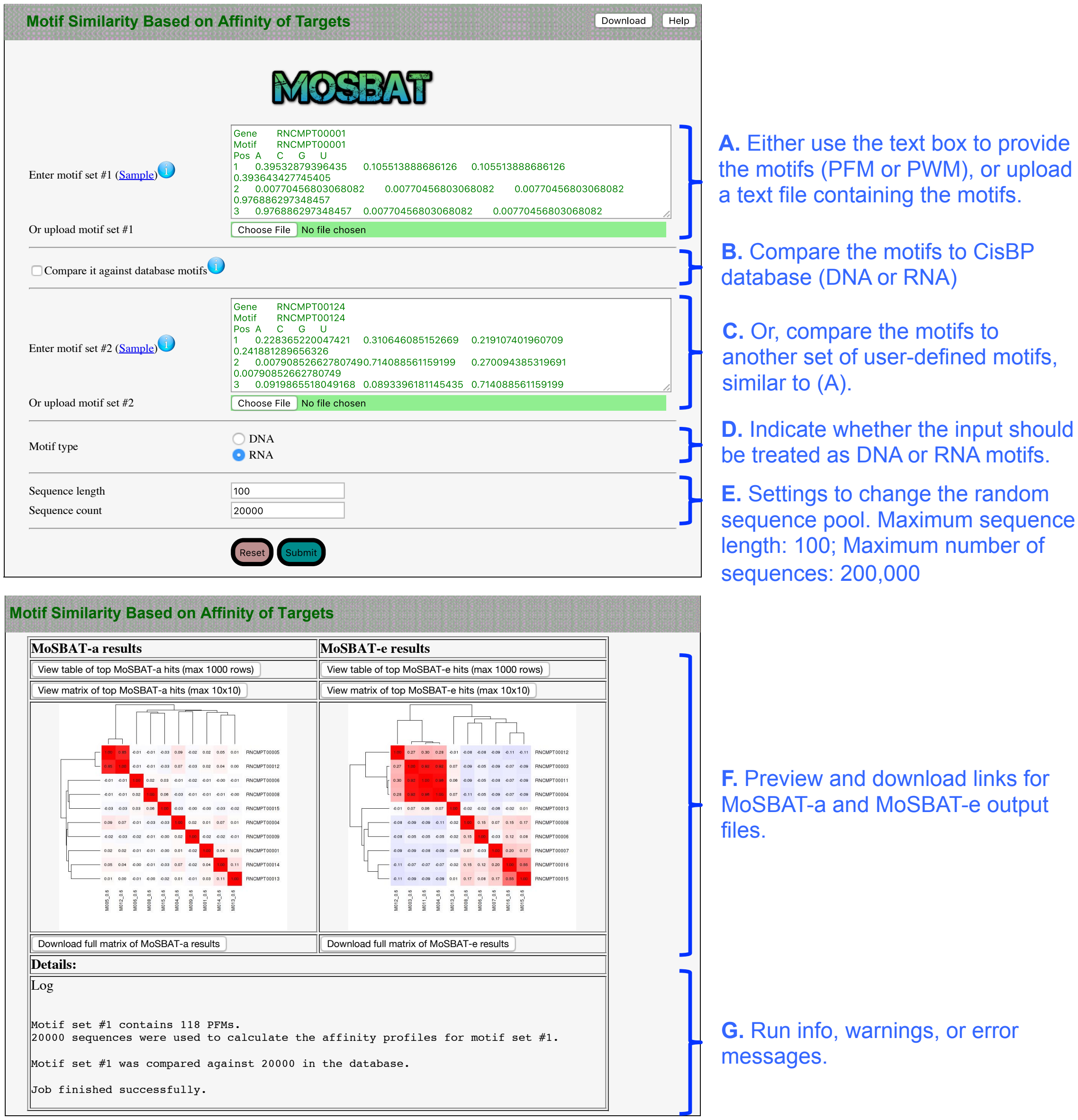
Foat, B.C., Morozov, A.V. and Bussemaker, H.J. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* 2006;22(14):e141-149.

Mahony, S. and Benos, P.V. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res* 2007;35(Web Server issue):W253-258.

Sandelin, A. and Wasserman, W.W. Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J Mol Biol* 2004;338(2):207-215.

Zhao, Y. and Stormo, G.D. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat Biotechnol* 2011;29(6):480-483.

SUPPLEMENTARY FIGURES



Motif Similarity Based on Affinity of Targets

Enter motif set #1 (Sample)

```
Gene	RNCMPT00001
Motif	RNCMPT00001
Pos	A	C	G	U
1	0.39532879396435	0.105513888686126	0.105513888686126
0.393643427745405
2	0.00770456803068082	0.00770456803068082	0.00770456803068082
0.976886297348457
3	0.976886297348457	0.00770456803068082	0.00770456803068082
```

Or upload motif set #1 Choose File No file chosen

□ Compare it against database motifs

Enter motif set #2 (Sample)

```
Gene	RNCMPT00124
Motif	RNCMPT00124
Pos	A	C	G	U
1	0.228365220047421	0.310646085152669	0.219107401960709
0.241881289656326
2	0.00790852662780749	0.714088561159199	0.270094385319691
0.00790852662780749
3	0.0919865518049168	0.0893396181145435	0.714088561159199
```

Or upload motif set #2 Choose File No file chosen

Motif type ○ DNA ● RNA

Sequence length: 100

Sequence count: 20000

A. Either use the text box to provide the motifs (PFM or PWM), or upload a text file containing the motifs.

B. Compare the motifs to CisBP database (DNA or RNA)

C. Or, compare the motifs to another set of user-defined motifs, similar to (A).

D. Indicate whether the input should be treated as DNA or RNA motifs.

E. Settings to change the random sequence pool. Maximum sequence length: 100; Maximum number of sequences: 200,000

Motif Similarity Based on Affinity of Targets

MoSBAT-a results

View table of top MoSBAT-a hits (max 1000 rows)
View matrix of top MoSBAT-a hits (max 10x10)

	M005_0.6	M012_0.6	M006_0.6	M008_0.6	M015_0.6	M004_0.6	M009_0.6	M001_0.6	M014_0.6	M013_0.6
RNCMPT00005	1.00	0.85	-0.01	-0.01	-0.03	0.09	-0.02	0.02	0.05	0.01
RNCMPT00012	0.85	1.00	-0.01	-0.01	-0.03	0.07	-0.03	0.02	0.04	0.00
RNCMPT00006	-0.01	-0.01	1.00	0.02	0.03	-0.01	-0.02	-0.01	-0.00	-0.01
RNCMPT00008	-0.01	-0.01	0.02	1.00	0.06	-0.03	-0.01	-0.01	-0.01	-0.00
RNCMPT00015	-0.03	-0.03	0.03	0.06	1.00	-0.03	-0.00	-0.00	-0.03	-0.02
RNCMPT00004	0.09	0.07	-0.01	-0.03	-0.03	1.00	0.02	0.01	0.07	0.01
RNCMPT00009	-0.02	-0.03	-0.02	-0.01	-0.00	0.02	1.00	-0.02	-0.02	-0.01
RNCMPT00001	0.02	0.02	-0.01	-0.01	-0.00	0.01	-0.02	1.00	0.04	0.03
RNCMPT00014	0.05	0.04	-0.00	-0.01	-0.03	0.07	-0.02	0.04	1.00	0.11
RNCMPT00013	0.01	0.00	-0.01	-0.00	-0.02	0.01	-0.01	0.03	0.11	1.00

MoSBAT-e results

View table of top MoSBAT-e hits (max 1000 rows)
View matrix of top MoSBAT-e hits (max 10x10)

	M012_0.6	M003_0.6	M011_0.6	M004_0.6	M013_0.6	M008_0.6	M006_0.6	M007_0.6	M016_0.6	M015_0.6
RNCMPT00012	1.00	0.27	0.30	0.28	-0.01	-0.08	-0.08	-0.09	-0.11	-0.11
RNCMPT00003	0.27	1.00	0.92	0.92	0.07	-0.09	-0.05	-0.09	-0.07	-0.09
RNCMPT00011	0.30	0.92	1.00	0.98	0.06	-0.09	-0.05	-0.08	-0.07	-0.09
RNCMPT00004	0.28	0.92	0.98	1.00	0.07	-0.11	-0.05	-0.09	-0.07	-0.09
RNCMPT00013	-0.01	0.07	0.06	0.07	1.00	-0.02	-0.02	-0.06	-0.02	0.01
RNCMPT00008	-0.08	-0.09	-0.09	-0.11	-0.02	1.00	0.15	0.07	0.15	0.17
RNCMPT00006	-0.08	-0.05	-0.05	-0.05	-0.02	0.15	1.00	-0.03	0.12	0.08
RNCMPT00007	-0.09	-0.09	-0.08	-0.09	-0.06	0.07	-0.03	1.00	0.20	0.17
RNCMPT00016	-0.11	-0.07	-0.07	-0.07	-0.02	0.15	0.12	0.20	1.00	0.55
RNCMPT00015	-0.11	-0.09	-0.09	-0.09	0.01	0.17	0.08	0.17	0.55	1.00

Details:

Log

```
Motif set #1 contains 118 PFMs.
20000 sequences were used to calculate the affinity profiles for motif set #1.

Motif set #1 was compared against 20000 in the database.

Job finished successfully.
```

F. Preview and download links for MoSBAT-a and MoSBAT-e output files.

G. Run info, warnings, or error messages.

Supplementary Figure 1. The input (**top**) and output (**bottom**) interface for MoSBAT web-server.

A

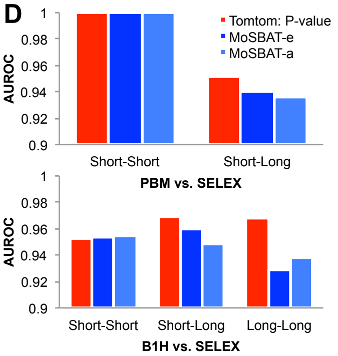
B1H vs. PBM	Nuclear receptor	Homeodomain	Average Rank
MoSBAT-e	0.72	0.81	1.5
Tomtom: P-value	0.71	0.81	2.5
PWMClus: PCC	0.63	0.84	4.0
MoSBAT-a	0.65	0.80	3.5
SSTAT: balanced - max	0.65	0.64	5.0
SSTAT: type 1 - sum	0.65	0.55	6.0
SSTAT: balanced - sum	0.65	0.55	6.5
PWMClus: Euclidean distance	0.58	0.73	7.5
STAMP	0.61	0.71	7.5
SSTAT: type 1 - max	0.63	0.58	7.5

B

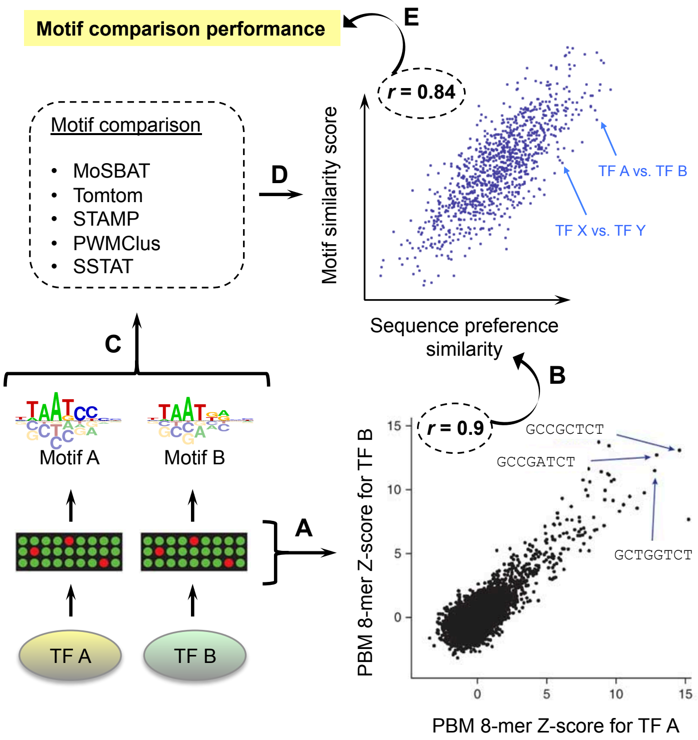
PBM vs. SELEX	Homeodomain,POU	GCM	Forkhead	Sox	Nuclear receptor	Homeodomain	RFX	Ets	bHLH	C2H2 ZF	bZIP	Average Rank
Tomtom: P-value	0.44	0.48	0.52	0.62	0.72	0.81	0.67	0.75	0.89	0.98	0.92	3.5
MoSBAT-e	0.38	0.48	0.52	0.69	0.76	0.77	0.56	0.83	0.94	0.92	0.93	3.5
MoSBAT-a	0.50	0.44	0.52	0.65	0.68	0.72	0.78	0.73	0.90	0.89	0.91	3.9
PWMClus: PCC	0.44	0.52	0.52	0.60	0.66	0.71	0.67	0.69	0.92	0.87	0.91	4.7
SSTAT: balanced - sum	0.50	0.60	0.54	0.62	0.65	0.55	0.67	0.81	0.66	0.73	0.89	5.0
SSTAT: balanced - max	0.50	0.52	0.54	0.55	0.61	0.61	0.78	0.81	0.64	0.78	0.84	5.1
SSTAT: type 1 - sum	0.50	0.48	0.51	0.53	0.61	0.55	0.83	0.82	0.66	0.74	0.74	6.1
PWMClus: Euclidean distance	0.50	0.48	0.51	0.53	0.61	0.63	0.67	0.69	0.83	0.73	0.89	6.3
STAMP	0.44	0.40	0.49	0.60	0.66	0.74	0.67	0.48	0.86	0.85	0.87	6.5
SSTAT: type 1 - max	0.50	0.44	0.52	0.52	0.60	0.56	0.78	0.79	0.73	0.71	0.74	7.1

C

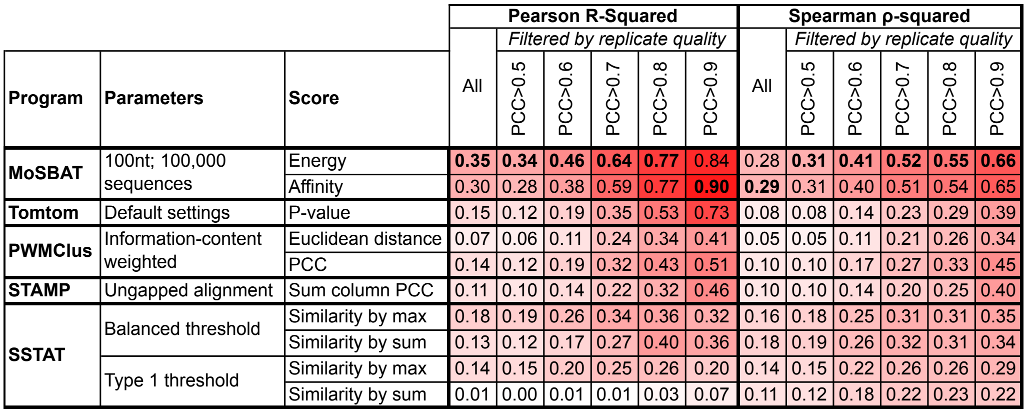
B1H vs. SELEX	T-box	Sox	Paired box	GATA	Forkhead	Ets	Nuclear receptor	Homeodomain	bHLH	bZIP	MADF	C2H2 ZF	Average Rank
MoSBAT-e	0.56	0.51	0.65	0.58	0.62	0.65	0.78	0.73	0.84	0.83	0.97	0.95	2.7
Tomtom: P-value	0.50	0.56	0.56	0.55	0.63	0.66	0.73	0.78	0.83	0.83	0.81	0.96	3.8
MoSBAT-a	0.53	0.56	0.57	0.58	0.66	0.62	0.75	0.72	0.83	0.79	0.84	0.95	4.3
STAMP	0.39	0.57	0.60	0.51	0.57	0.70	0.71	0.75	0.77	0.83	0.91	0.97	4.4
Tomtom: Q-value	0.50	0.56	0.57	0.55	0.56	0.63	0.73	0.72	0.82	0.81	0.47	0.92	5.3
SSTAT: balanced - max	0.42	0.49	0.56	0.57	0.69	0.62	0.61	0.69	0.80	0.81	1.00	0.93	5.2
SSTAT: balanced - sum	0.56	0.49	0.50	0.54	0.62	0.59	0.63	0.59	0.76	0.84	1.00	0.87	6.3
PWMClus: Euclidean distance	0.42	0.46	0.49	0.57	0.50	0.65	0.53	0.66	0.79	0.80	0.97	0.91	6.8
SSTAT: type 1 - sum	0.43	0.47	0.43	0.51	0.67	0.64	0.67	0.61	0.73	0.77	0.92	0.84	7.6
SSTAT: type 1 - max	0.40	0.47	0.39	0.48	0.65	0.64	0.68	0.65	0.75	0.80	0.91	0.87	7.7



Supplementary Figure 2. Performance of MoSBAT and four other methods for labeling *in vitro* motifs based on comparison with motifs from other *in vitro* assays (similar to Supplementary Figure 4). (**A-C**) AUROC values for correctly labeling the motifs from one assay based on comparison to motifs from a different assay. TF structural classes that are examined for different assay types are selected based on availability of data. (**D**) Comparison of performance of MoSBAT with Tomtom, stratified based on motif length (short: ≤14 nucleotides wide; long: >14 nucleotides wide). Only motifs from the C2H2-ZF class are included in this analysis, as TFs of other structural classes rarely have long motifs (Weirauch *et al.*, Cell 2014, 158:1431-1443; Najafabadi *et al.*, Nat Biotechnol 2015, 33:555-562). There were not enough examples to include a long-long category for PBM vs. SELEX comparison. The scale of the y-axis is between 0.9–1.0 to magnify the differences.

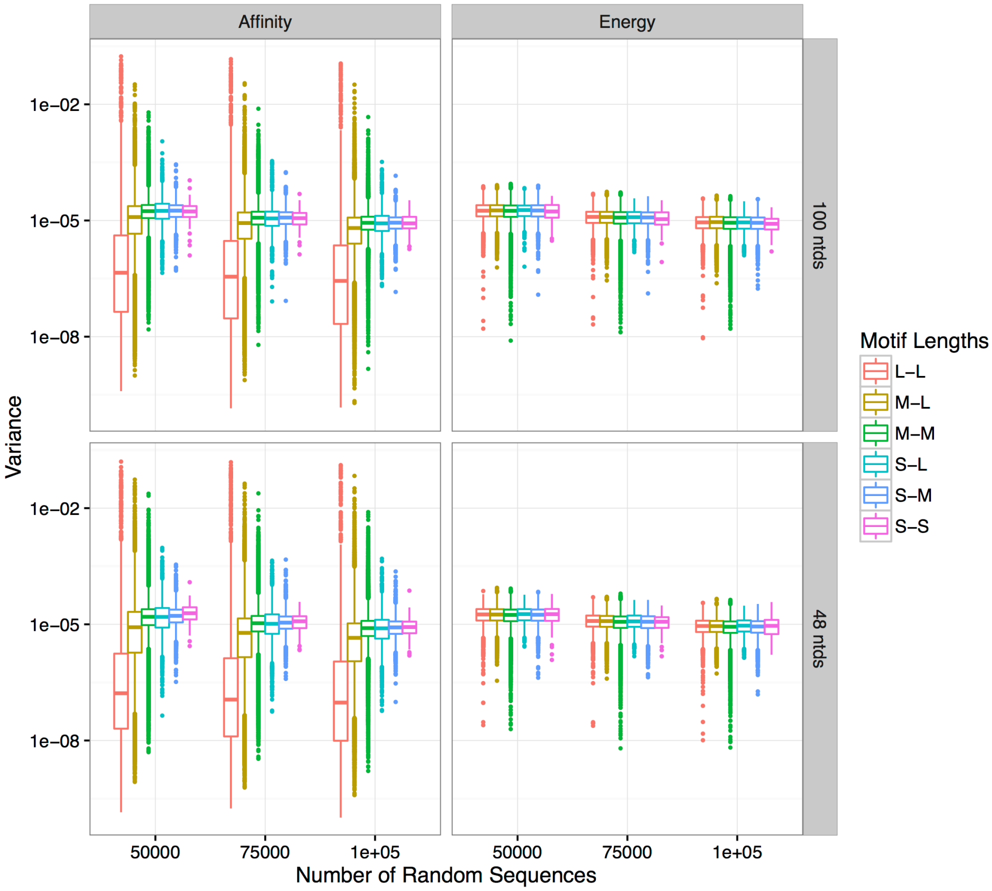


Supplementary Figure 3. Schematic presentation of the benchmarking workflow for comparison of motif similarity scores to TF sequence preference similarity. For each pair of TFs, (**A**) we first calculated the Pearson coefficient of the PBM Z-scores of 32,896 unique 8-mers (Berger *et al.*, Nat Biotechnol 2006, 24:1429-1435). The Z-score reflects the binding affinity of a TF for each DNA 8-mer. Therefore, the calculated Pearson coefficient for the two PBMs (**B**) serves as a direct measure of similarity of the sequence preferences of the two TFs. In parallel (**C**), we calculated the pairwise similarity of motifs that were derived from the same set of PBMs, using MoSBAT or each of other four motif comparison tools. The motifs were derived using a uniform processing pipeline for all PBMs (Weirauch *et al.*, Cell 2014, 158:1431-1443). We then compared the motif similarity scores with the 8-mer–based TF preference similarities (**D**), in order to obtain a measure of performance (**E**) for each of the motif comparison tools. Figure 1B and Supplementary Figure 3 show a summary of the obtained performance values.



			Pearson R-Squared						Spearman ρ-squared					
				Filtered by replicate quality						*Filtered by replicate quality*				
Program	**Parameters**	**Score**	All	PCC>0.5	PCC>0.6	PCC>0.7	PCC>0.8	PCC>0.9	All	PCC>0.5	PCC>0.6	PCC>0.7	PCC>0.8	PCC>0.9
MoSBAT	100nt; 100,000 sequences	Energy	**0.35**	**0.34**	**0.46**	**0.64**	**0.77**	0.84	0.28	**0.31**	**0.41**	**0.52**	**0.55**	**0.66**
		Affinity	0.30	0.28	0.38	0.59	0.77	**0.90**	**0.29**	0.31	0.40	0.51	0.54	0.65
Tomtom	Default settings	P-value	0.15	0.12	0.19	0.35	0.53	0.73	0.08	0.08	0.14	0.23	0.29	0.39
PWMClus	Information-content weighted	Euclidean distance	0.07	0.06	0.11	0.24	0.34	0.41	0.05	0.05	0.11	0.21	0.26	0.34
		PCC	0.14	0.12	0.19	0.32	0.43	0.51	0.10	0.10	0.17	0.27	0.33	0.45
STAMP	Ungapped alignment	Sum column PCC	0.11	0.10	0.14	0.22	0.32	0.46	0.10	0.10	0.14	0.20	0.25	0.40
SSTAT	Balanced threshold	Similarity by max	0.18	0.19	0.26	0.34	0.36	0.32	0.16	0.18	0.25	0.31	0.31	0.35
		Similarity by sum	0.13	0.12	0.17	0.27	0.40	0.36	0.18	0.19	0.26	0.32	0.31	0.34
	Type 1 threshold	Similarity by max	0.14	0.15	0.20	0.25	0.26	0.20	0.14	0.15	0.22	0.26	0.26	0.29
		Similarity by sum	0.01	0.00	0.01	0.01	0.03	0.07	0.11	0.12	0.18	0.22	0.23	0.22

Supplementary Figure 4. Comparison of performance of MoSBAT and four other methods, using PBM assays with varying degrees of reproducibility (PCC: Pearson correlation coefficient of Z-scores between ME/HK sets of probes; see Weirauch *et al.*, Nat Biotechnol 2013, 31:126-134). The values in the table represent the correlation of motif similarity score vs. PBM similarity (see Supplementary Figure 3 for details).



Supplementary Figure 5. Distribution of variance in motif similarity scores. Motif similarity scores were calculated using MoSBAT-a (affinity) or MoSBAT-e (energy) between ~100,000 PBM and HT-SELEX motif pairs on 10 sets of random sequences for two sequence lengths (48 or 100 nucleotides, 50 %GC). The variance of scores across the 10 replicates was calculated for each motif, and the distribution of variance was plotted. Variances are grouped based on motif lengths: short (S; <8 nt), medium (M, 8–12 nt), and long (L, ≥13 nt).