

Supplemental Information

A novel *k*-mer set memory (KSM) motif representation improves regulatory variant prediction

Yuchun Guo, Kevin Tian, Haoyang Zeng, Xiaoyun Guo, David K Gifford

MIT, Computer Science and Artificial Intelligence Laboratory, Cambridge, MA

Supplemental Table S1. Related to Figure 1

Number of component k -mers (length= k , gap=0~4) in the primary KSM motifs trained from top n sequences of representative ENCODE ChIP-seq datasets

Experiment	k	$n=5000$	$n=10000$	$n=30000$	$n=50000$
JunD_SnyderS_K562	8	639	1075	1786	2710
EBF1--SC-137065-PCR1x_Myers_GM12878	8	852	1314	2458	2960
YY1--SC-281-PCR1x_Myers_GM12878	8	758	756	1814	2586
Max-v041610.2_Myers_K562	9	1227	2502	6911	10908
CTCF_Crawford_K562	10	2455	3155	9452	12416
PU.1-PCR1x_Myers_GM12878	10	1929	5382	16937	26400

Supplemental Table S2. Related to Figure 1

Partial AUROC (fpr \leq 0.1) in predicting held-out ChIP-seq sequences using primary KSM motifs trained from top n sequences of representative ENCODE ChIP-seq datasets

Experiment	k	$n=5000$	$n=10000$	$n=30000$	$n=50000$
JunD_SnyderS_K562	8	0.0338	0.0367	0.0376	0.0383
EBF1--SC-137065-PCR1x_Myers_GM12878	8	0.0243	0.0253	0.0233	0.0160
YY1--SC-281-PCR1x_Myers_GM12878	8	0.0300	0.0330	0.0314	0.0304
Max-v041610.2_Myers_K562	9	0.0304	0.0309	0.0328	0.0321
CTCF_Crawford_K562	10	0.0456	0.0512	0.0554	0.0565
PU.1-PCR1x_Myers_GM12878	10	0.0654	0.0699	0.0726	0.0725

Supplemental Table S3. Related to Figure 2B

Number of experiments of which motifs are correctly re-discovered by running KMAC on various number of sequences.

# Seqs	1000	700	500	300	200	100	50	40	30
# Correct	178	171	168	168	159	152	139	132	126

Supplemental Table S4. Related to Figure 2B

Running time of five *de novo* motif discovery methods on top 1,000 sequences from four representative datasets

Motif discovery method	KMAC	Weeder2	HOMER	MEME-Chip	Slim
c-Myc_Crawford_K562	2m06	4m58	38m55	46m01	100m41
CTCF_Bernstein_GM12878	5m34	4m43	39m02	39m28	135m30
GABP-PCR2x_Myers_K562	4m35	4m35	44m53	40m11	102m46
NRSF-PCR2x_Myers_K562	5m45	5m02	40m45	44m07	103m19

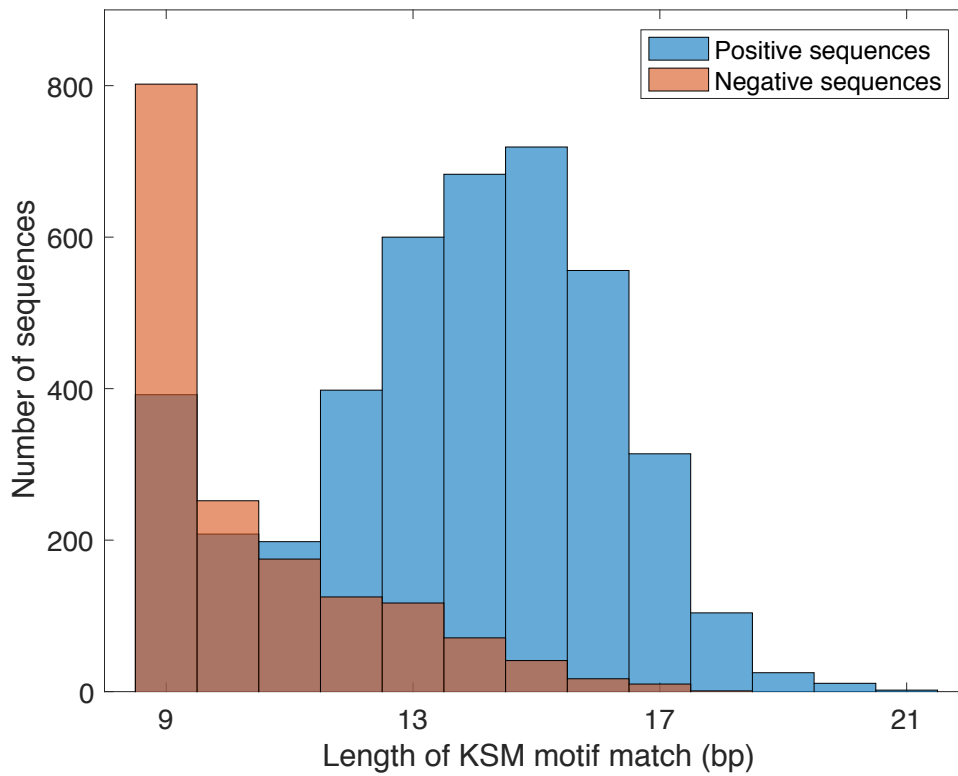
Supplemental Table S5. Related to Figure 3 and 4

Running time of four motif models for scanning motifs on 100,000 101bp sequences

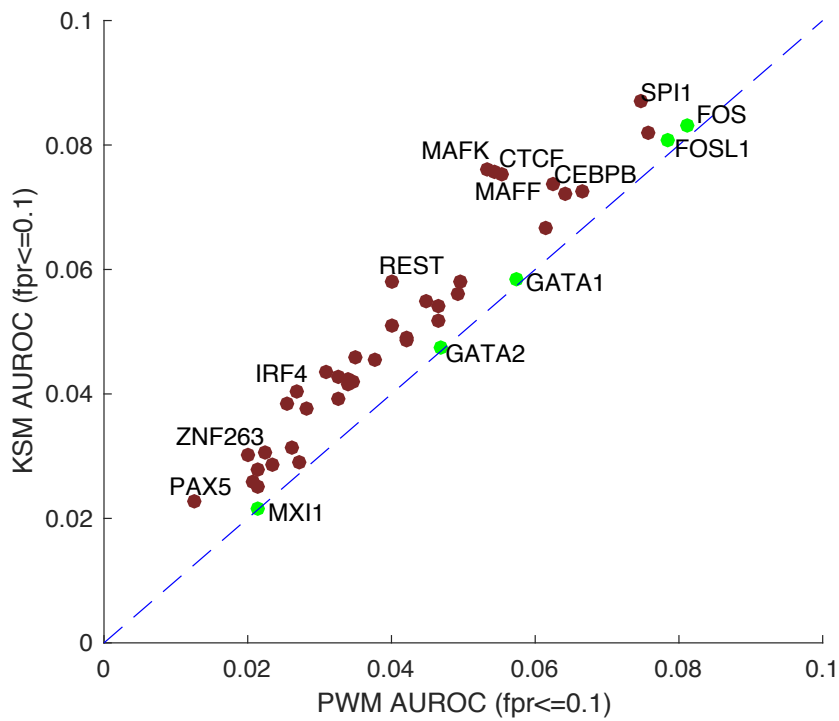
Motif model	KSM	PWM	Slim	TFFM
Motif scanning time	5.5s	2.1s	160.6s	563.9s

K-mers	Sequence hit bit string									Hit count
---ATGCAAAT	1	1	1	1	1	0	0	0	0	5
--TATGCAA	0	1	1	1	0	0	0	0	1	4
----TGCAAATG	0	0	1	0	0	1	1	0	0	3
K-mer Group 1	1	1	1	1	1	1	1	0	1	8

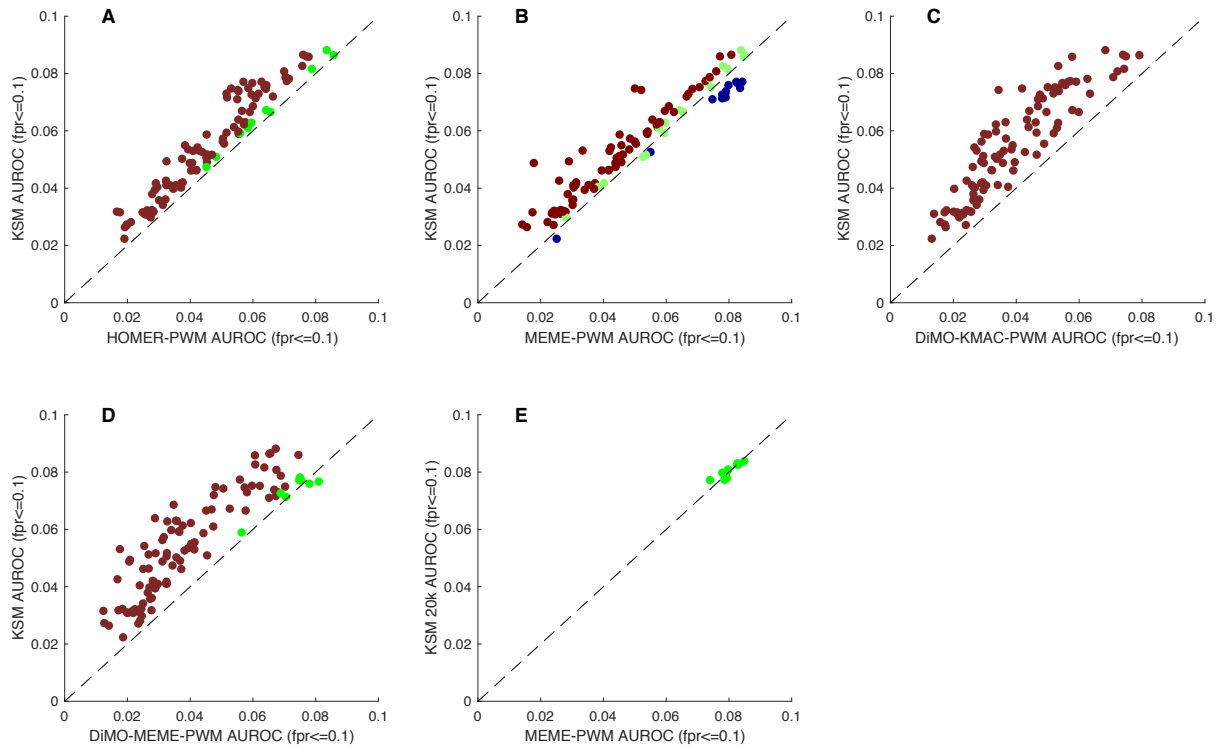
Supplemental Figure S1. Related to Figure 1B. The hit count of a *k*-mer group is computed by a union operation on sequence hit bit strings associated with the matched *k*-mers in the *k*-mer group. In this example, there are nine training sequences. The presence/absence of a *k*-mer in all nine training sequences is stored as a 9-bit bit string associated with the *k*-mer.



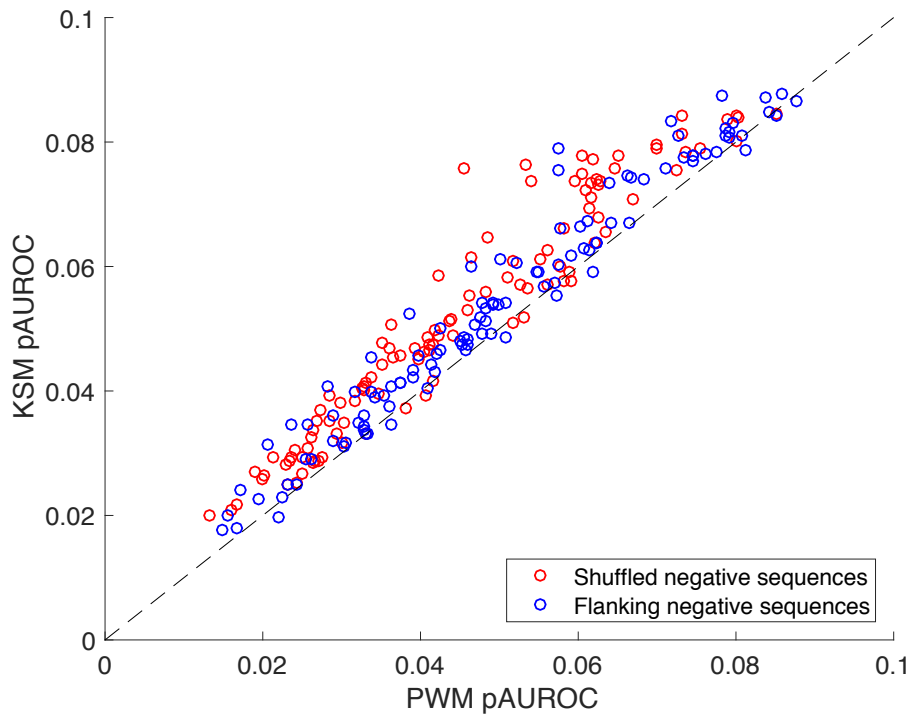
Supplemental Figure S2. Related to Figure 3B. KSM motif matches in positive sequences include more flanking sequences than those in negative sequences. Histograms showing the distribution of the length of GABP KSM motif matches in positive vs. negative test sequences.



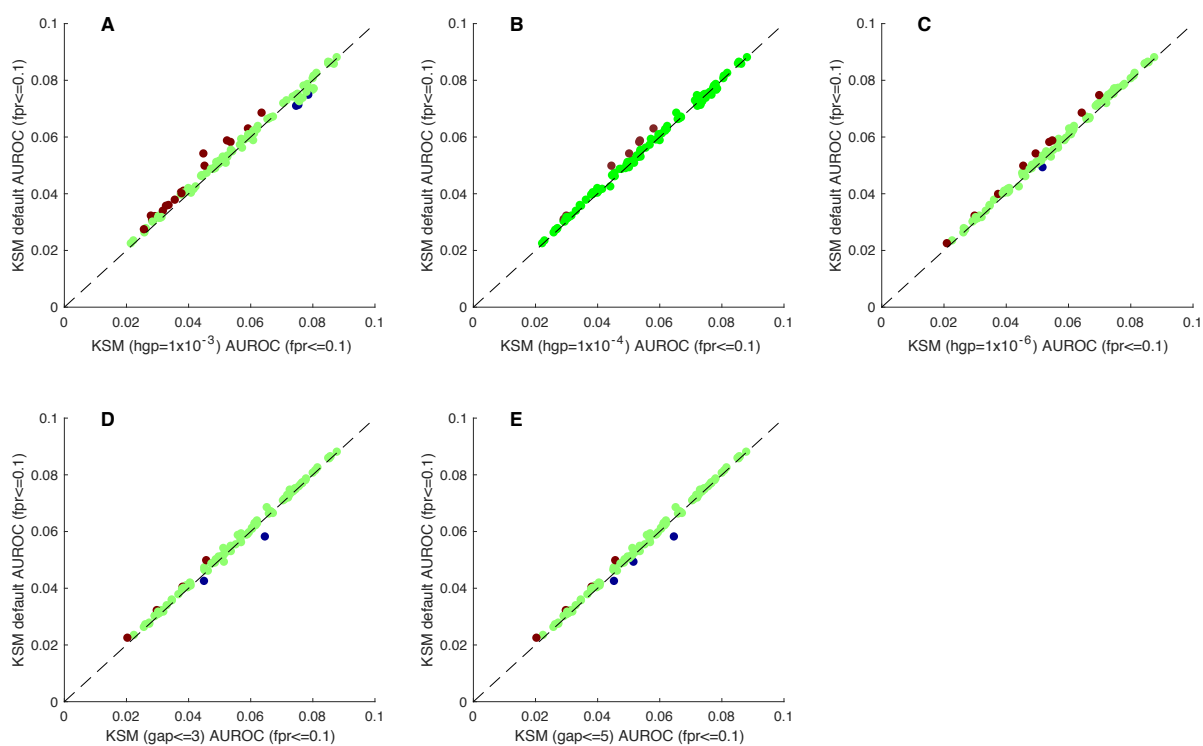
Supplemental Figure S3. Related to Figure 3C. Scatter plot comparing the mean partial AUROC (fpr ≤ 0.1) values of KSM and PWM for predicting ChIP-seq binding for 43 TFs.



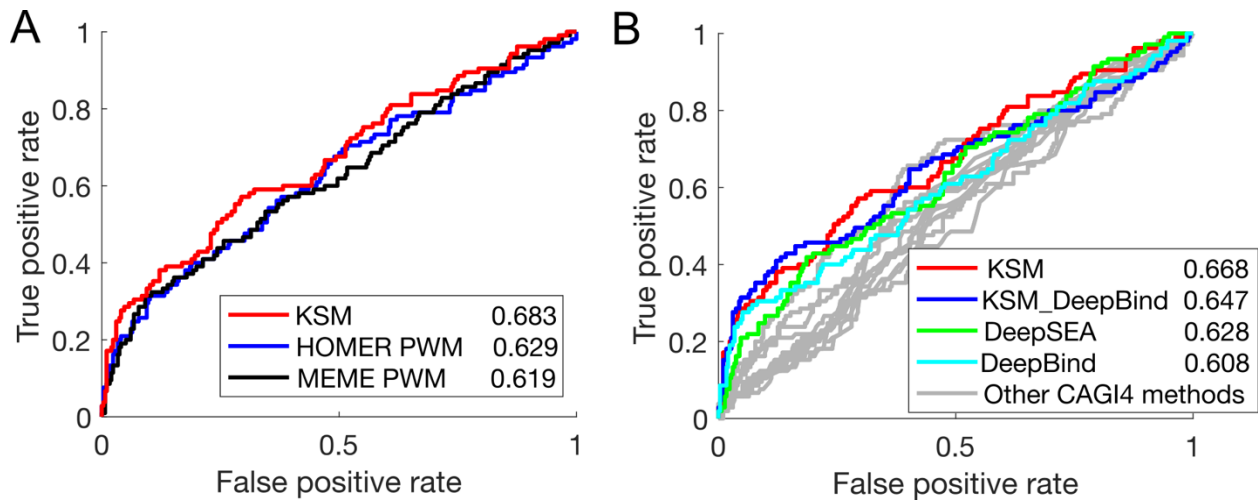
Supplemental Figure S4. Related to Figure 3C. Scatter plot comparing the mean partial AUROC (fpr ≤ 0.1) values of KSM and various PWMs for predicting ChIP-seq binding for 104 experiments. (A) KSM vs. HOMER PWM. (B) KSM vs. MEME PWM. (C) KSM vs. DiMO optimized KMAC PWM. (D) KSM vs. DiMO optimized MEME PWM. (E) KSM trained from 20,000 sequences vs. MEME PWM for 11 CTCF experiments. Each point represents a ChIP-seq experiment. Brown: the KSM performs better than the PWM; blue: the KSM performs worse; green: both representations perform similarly.



Supplemental Figure S5. Related to Figure 3C. KSM outperforms PWM in discriminating in vivo TF bound sequences from various types of negative sequences.



Supplemental Figure S6. Related to Figure 3C. Scatter plot comparing the accuracy values of default KSM (k -mer hgp cutoff = 1×10^{-5} , $\text{gap} \leq 4$) and KSMs with different parameters: (A) k -mer hgp cutoff = 1×10^{-3} , $\text{gap} \leq 4$; (B) k -mer hgp cutoff = 1×10^{-4} , $\text{gap} \leq 4$; (C) k -mer hgp cutoff = 1×10^{-5} , $\text{gap} \leq 4$; (D) k -mer hgp cutoff = 1×10^{-5} , $\text{gap} \leq 3$; (E) k -mer hgp cutoff = 1×10^{-5} , $\text{gap} \leq 5$. The performance for predicting ChIP-seq binding is evaluated as mean partial AUROC ($\text{fpr} \leq 0.1$) for 104 experiments. Each point represents a ChIP-seq experiment. Brown: the KSMs with default parameters perform better than the KSMs with other parameters; blue: the KSM with default parameters performs worse; green: both representations perform similarly.



Supplemental Figure S7. Related to Figure 6. KSMs predicts allele-specific differences in regulatory activity better than PWMs and deep learning derived features.

(A) ROC performance of KSM and PWM motif representation in predicting differential regulatory activities of eQTL alleles. The numeric values in the legend are the AUROC values. (B) Similar to (A), KSM, DeepSEA derived features and other CAGI 4 open challenge methods.