Supplementary Material

Using DNase digestion data to accurately identify transcription factor binding sites

KAIXUAN LUO[1] and ALEXANDER J. HARTEMINK[1,2]

[1]Program in Computational Biology and Bioinformatics,
[2]Department of Computer Science,
Duke University, Durham, NC 27708, USA

E-mail: kaixuan.luo@duke.edu, amink@cs.duke.edu

Supplement Table 1. Average model performance in 20 yeast and six human TFs using MILLIPEDE gold standard.

Each row refers to a different method. Columns are auROC, auPR, sensitivity at 1% FPR, precision at 1% FPR in yeast and human.

	Yeast TF mean				Human TF mean			
Models	**auROC (%)**	**auPR (%)**	**Sens. (%)**	**Prec. (%)**	**auROC (%)**	**auPR (%)**	**Sens. (%)**	**Prec. (%)**
M24	NA	NA	NA	NA	97.4	57.0	63.9	50.4
M12	93.4	28.1	33.3	31.5	97.8	57.2	64.6	50.6
M11	93.4	28.2	33.4	31.6	97.8	57.2	61.8	50.5
M5	94.2	29.1	34.5	32.6	97.6	56.1	61.5	50.3
M3	93.8	28.3	33.0	32.1	97.6	55.6	58.8	50.0
M2	93.8	27.4	32.2	31.4	97.6	55.0	59.1	50.2
M1	93.9	27.5	32.4	31.5	97.6	55.0	59.0	50.1
M1 w/o PWM	88.9	12.3	10.5	12.2	96.3	50.0	45.4	46.6
PWM only	86.2	16.7	16.5	22.1	71.2	24.3	14.1	31.4
CENTIPEDE w/ shrinkage	73.6	12.8	14.2	16.4	97.3	48.3	51.2	47.7
CENTIPEDE w/o shrinkage	55.3	11.3	15.4	15.0	94.3	49.0	57.0	49.2

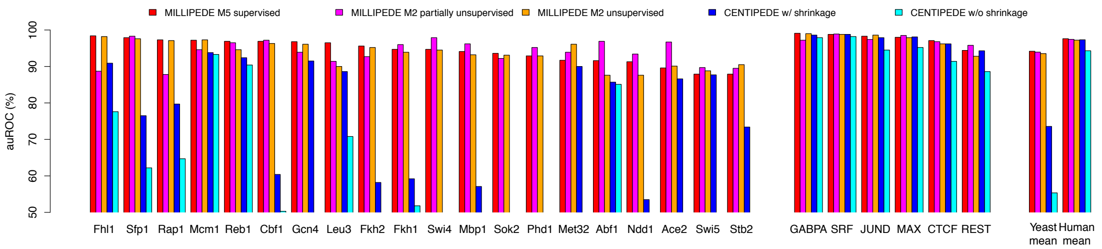
Supplement Table 2. Average model performance in six human TFs using MILLIPEDE and CENTIPEDE gold standard.

Each row refers to a different method. Columns are auROC, auPRC, sensitivity at 1% FPR, precision at 1% FPR using MILLIPEDE and CENTIPEDE gold standard.

	MILLIPEDE gold standard				CENTIPEDE gold standard			
Models	**auROC (%)**	**auPR (%)**	**Sens. (%)**	**Prec. (%)**	**auROC (%)**	**auPR (%)**	**Sens. (%)**	**Prec. (%)**
M24	97.4	57.0	63.9	50.4	98.4	73.8	80.4	60.4
M12	97.8	57.2	64.6	50.6	98.8	75.2	84.7	60.9
M11	97.8	57.2	61.8	50.5	98.7	75.1	84.3	60.7
M5	97.6	56.1	61.5	50.3	98.6	74.1	82.2	60.8
M3	97.6	55.6	58.8	50.0	98.6	74.0	80.0	60.8
M2	97.6	55.0	59.1	50.2	98.6	73.9	80.8	60.8
M1	97.6	55.0	59.0	50.1	98.6	73.8	80.3	60.8
M1 w/o PWM	96.3	50.0	45.4	46.6	97.1	70.4	72.2	59.3
PWM only	71.2	24.3	14.1	31.4	73.0	30.5	19.2	35.1
CENTIPEDE w/ shrinkage	97.3	48.3	51.2	47.7	98.0	69.8	81.8	60.2
CENTIPEDE w/o shrinkage	94.3	49.0	57.0	49.2	86.5	60.6	68.7	55.2

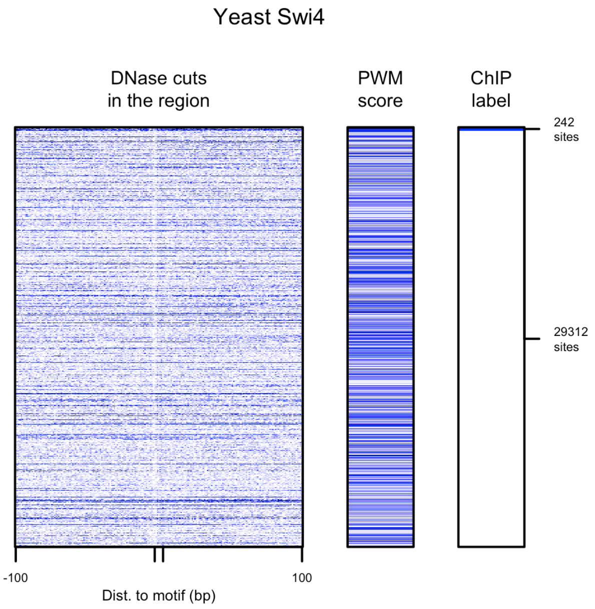
Supplement Figure 1. Area under the ROC curve for 20 yeast and six human TFs.

Red bars are MILLIPEDE model M5 run in a supervised mode, magenta bars are M2 run in a partially unsupervised mode (trained by Reb1 in yeast, or REST in human), orange bars are M2 run in a completely unsupervised mode, blue bars are CENTIPEDE with shrinkage, cyan bars are CENTIPEDE without shrinkage. Bars start at 50% since that represents random performance for an ROC curve; values below 50 are just not shown. The 20 yeast TFs are listed before the six human TFs; within each organism, the TFs are sorted such that the red bars decrease in height.



Supplement Figure 2. DNase digestion data for Swi4 candidate binding sites in yeast.

DNase data can exhibit systematic artifacts such as sequence-dependent digestion bias. The figure shows the digestion patterns for Swi4 candidate binding sites, whose consensus binding sequence is CGCGAAA.


Yeast Swi4
DNase cuts in the region
PWM score
ChIP label
242 sites
29312 sites
Dist. to motif (bp)