Published online

Nucleic Acids Research, 0000, Vol. 00, No. 00 ***1–20***
doi:10.1093/nar/gkn000

Supporting Material for: COUGER – CO-factors associated with Uniquely-bound GEnomic Regions

Alina Munteanu [1,2], Uwe Ohler [2,3], Raluca Gordân [3*]

[1]Faculty of Computer Science, Alexandru I. Cuza University, Iasi 700483, Romania, [2]Berlin Institute for Medical Systems Biology, Max Delbruck Center, 13125 Berlin, Germany and [3]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27708, USA

MATERIALS AND METHODS

Peak calling pipeline

Our approach compares bound regions (ChIP-seq peaks) derived from multiple experiments, possibly conducted in different laboratories. For this reason, we wanted to use a uniform processing pipeline for the peak calling step. With regards to the actual peak calling algorithm, we reduced the candidate list to MACS (1) and SPP (2). The selection was based on their wide usage as well as their reviewed performance. Both MACS and SPP have a high number of identified peaks and good sensitivity, but most importantly, they have a very good positional accuracy (estimated as the distance between the summit and the closest high confidence occurrence of a motif) (3).

We ran the IDR (Irreproducible Discovery Rate) pipeline (4) with these two peak calling algorithms: MACS and SPP. For MACS we used two variants of version 2.0.10: one with the default estimation of the shift size, and another variant with the size of the shift previously estimated by SPP. Both variants use the parameters needed for relaxed thresholds regarding peak confidence (-p 1e-3 –to-large). In the case of the SPP peak caller, we used the 1.10 version with the parameters recommended by IDR usage (-npeak=300000 -savr -savp -rf).

We compared the peaks obtained by us, using these three algorithms (with IDR), and two other sets of peaks available from ENCODE (5): the peaks reported in the ENCODE database by the laboratory that performed the experiment, and the peaks reported by the ENCODE Analysis Working Group (AWG) on March 2012 Freeze, that are based on IDR with SPP.

We analyzed the "quality" of a peak caller by scanning the resulted peaks for 2 consecutive 8-mers that have the PBM E-score greater than 0.4, and then plotting the percent of peaks that had at least one hit. This measure is similar to a motif scan, but uses PBM data instead of PWM data, and is highly informative for our particular setting. Moreover, because we use a fixed region centered at the summit of the ChIP-seq peaks, we considered not only the full length of the peaks in this analysis, but also the 200 bp and 300 bp region centered at the summit.

Fig. S7 shows the results of this analysis for c-Myc transcription factor. We can see that a large fraction of peaks contain putative TF binding sites (Fig. S7A), which indicates that the ChIP dataset is of high quality. Also, putative c-Myc binding sites are enriched in ChIP-seq peaks relative to open chromatin regions (i.e., DNase-seq peaks) – Fig. S7B. The variance between the two subplots of Fig. S7 is due to the length distributions of the peaks identified by different peak callers. The default ENCODE peaks are quite long in general, having up to 10,000 bp, although the majority of them have less than 1,000 bp. The SPP peak caller has the opposite behavior, and the peaks are short, usually between 80 and 350 bp, with the vast majority of peaks having around 300 bp in length. MACS2 outputs more balanced peaks, in terms of their lengths, varying between 100 and 2000 bp in length, with a large of number of peaks having between 200 and 450 bp. Therefore the analysis that considers fixed length region of peaks is more reliable, while the analysis that considers full peaks reflects mostly their length.

We analyzed the results obtained in this setting for several TFs, and we concluded that the best option for peak calling is to use MACS, but to estimate the shift size with SPP.

*To whom correspondence should be addressed. Tel: +1 919 684 9881; Fax: +1 919 668 0795; Email: raluca.gordan@duke.edu

Using IDR to select reproducible sequences for TFs and replicates

In our study, we used the IDR framework (4) to process the ChIP-Seq data from ENCODE (5) in the K562 cell line, because it is the current standard for reporting data generated by the ENCODE project (6). IDR (Irreproducible Discovery Rate) (4) is a measure of the reproducibility of findings from replicate experiments. The IDR algorithm separates signal from noise on a pair of ranked lists of ChIP-seq peaks (or other identifications) from different replicates, creating a curve which assesses the consistency and reproducibility of these peaks. The method provides stable thresholds for a list of consistency and reproducibility levels.

We applied the IDR pipeline for all paralogous TFs that we considered. More precisely, for each factor, we performed peak calling (with MACS) and ran IDR analysis on several pairs of ChIP-seq mapped reads sets:

- the actual pair of replicates, obtaining the number of peaks consistent between true replicates (N_t).
- the pair of pooled pseudo-replicates, obtaining the number of peaks consistent between pooled pseudo-replicates (N_p). We generated these pooled pseudo-replicates by first pooling the mapped reads from all replicates and then randomly splitting this set into two equal sets of reads.
- the two pairs of self-pseudo-replicates, each corresponding to an actual replicate, obtaining the number of self-consistent peaks for each replicate (N_1, N_2). We obtained these self-pseudo-replicates by randomly splitting the mapped reads of the considered replicate into two equal sets of reads.

Next, we analyzed the results and filtered out the TFs with IDR scores that did not follow the restrictions recommended by ENCODE (6):

$$0.5 < N_1/N_2 < 2 \quad \textbf{(1)}$$

$$N_p/N_t < 2 \quad \textbf{(2)}$$

Equation (**1**) guarantees that the number of self-consistent bound regions identified from one replicate is at least half and at most double than the number of self-consistent bound regions identified from the other replicate, while equation (**2**) ensures that the number of peaks consistent between replicates is at least 50% of the number of regions consistent between two pseudo-replicates generated by randomly partitioning available reads from all replicates.

We discarded four ChIP-seq datasets (corresponding to STAT1 and STAT2 with IFNa6h and IFNa30 treatments) with IDR scores outside the limits (see Table S11). The remaining datasets represented 31 TFs: ATF1, ATF3, BHLHE40, CEBPB, CEBPD, CFOS, CJUN, CJUN*, CMYC, CMYC*, E2F4, E2F6, EFOS, EJUNB, EJUND, ELF1, ETS1, GABPA, GATA1, GATA2, JUND, MAFF, MAFK, MAX, MAX*, MXI1, NFYA, NFYB, SP1, SP2 and TAL1 (* indicates IgG control), in 20 pairs of paralogous TFs.

For each considered TF, we used the IDR cutoff (N_t) to select only the peaks that are reproducible between replicate experiments. This set of high confidence peaks were then used as input in the comparisons between paralogous TFs.

For the comparisons between replicate experiments, we used N_1 and N_2 thresholds to determine self-consistent peaks for each replicate. Hence we took into account only the top N_1 peaks derived from the first replicate and the top N_2 peaks derived from the second replicate.

Randomization procedure

For each randomized run, we started with the same sequences that were used in the normal run for the two classes. Before running **COUGER**, we pooled together these selected sequences (specific to each TF) and then randomly split them in two new files. The random class files were then used as input for both the run with PWM features (results in Table S3) and the run with PBM features (results in Table S4). We repeated the proccess five times, including the random splitting part, for a selected number of pairs of TFs (results in Table S5).

Calculation of accuracy, sensitivity, specificity, and precision

We are using supervised classification algorithms, and the class labels are known. Thus, we can compute the number of true/false positive and true/false negative. The accuracy, sensitivity, specificity, and precision are computed in a 5-fold cross-validation setting. Therefore, we train the algorithm on 4/5 of the sequences specific to TF1 and TF2, and then we predict the class for the remaining set (1/5 sequences), for which the true positives and true negatives are available. Importantly, we always test the classification models on sequences that were not used for training.

RESULTS

Stability of the results as the number of sequences decreases

We tested the influence of class size reduction in the case presented in Figure 2 of the main text, namelly Fos versus JunD with PBM features. Each of the TFs were represented in classification by 1016 sequences. Besides considering the whole set of sequences, we ran **COUGER** for this pair of paralogous TFs with maximum 800, 500 and 300 sequences in each class. The classification performance was very similar among different sized subsets (Table S13), the median accuracy having even a small increase with the decrease of number of considered sequences (91.13%, 91.56%, 94% and 95.83% for 1016, 800, 500, and 300 sequences per class, respectively). This improvement is likely due to the fact that the original sequences are representative for the TFs, and that the subsets are not chosen at random, but are sorted by their p-value. The different identified sets of NMIFS10 putative cofactors (first 10 features selected by normalized mutual information feature selection) are presented in Table S14.

Results obtained using PBM versus PWM features

In general, the results reported by COUGER when using PBM versus PWM features are similar, but not identical. For example, Tables S12 and S15 contain the

top 10 NMIFS-selected features when comparing the unique binding regions of Fos versus JunD. Factors from the ATF/CREB subfamily were identified by both methods (PBM feature "Atf1_3026_contig8mers_v1" and PWM feature "CREB1_MA0018.2_jaspar"). GATA factors, known to interact with proteins from the AP-1 subfamily, were also identified by both methods, as well as homeobox factors (e.g. Lhx6, and Hoxc9). As expected, though, the results are not identical. Each approach has its own advantages and disadvantages, as described in the main text. In general, we found that using PBM features helps us identify more putative cofactors that are supported by literature.

SUPPLEMENTARY FIGURES AND TABLES

(A)

(B) ...GCGC **TCCCCACGTGCGCGCTCACACGTGCATCGGCTACCCATGTGCCTC** GGGCT ...

10-mer	PWM score
TCCCCACGTG	-7.0013
CCCCACGTGC	-23.4061
CCCACGTGCG	**7.7842**
CCACGTGCGC	-30.0300
CACGTGCGCG	-0.6539
ACGTGCGCGC	-27.9614
CGTGCGCGCT	-4.0158
GTGCGCGCTC	-40.9659
TGCGCGCTCA	-0.7386
GCGCGCTCAC	-29.8710
CGCGCTCACA	-13.2007
GCGCTCACAC	-29.5139
CGCTCACACG	-8.9880
GCTCACACGT	-31.1448
CTCACACGTG	-2.2240
TCACACGTGC	-26.7710
CACACGTGCA	**5.6353**
ACACGTGCAT	-33.8801
CACGTGCATC	-9.1585
ACGTGCATCG	-20.8069
CGTGCATCGG	-10.1774
GTGCATCGGC	-29.7659
TGCATCGGCT	-9.2720
GCATCGGCTA	-15.0440
CATCGGCTAC	-24.3472
ATCGGCTACC	-18.3647
TCGGCTACCC	-20.7391
CGGCTACCCA	-19.8933
GGCTACCCAT	-23.5238
GCTACCCATG	-17.8316
CTACCCATGT	-16.9370
TACCCATGTG	-7.4740
ACCCATGTGC	-25.4189
CCCATGTGCC	**6.7985**
CCATGTGCCT	-25.9673
CATGTGCCTC	-15.7824

Figure S1. Illustration of **COUGER**'s feature derivation for a specific PWM and a specific DNA sequence (in **COUGER**, the DNA sequences are selected from ChIP-seq peaks unique to the paralogous TFs of interest). (A) Logo of a position weight matrix (PWM), in this case for TF c-Myc. (B) The computation of "MAX" and "TOP3AVG" features for a specific DNA sequence and the c-Myc PWM. A window of the same size as the PWM (in this case k=10) is slided across the sequence of interest, and a "PWM score" is computed for each window position. This score represents the log-likelihood ratio of that k-mer being generated by the PWM as opposed to a uniform background frequency model. Next, two numerical features are computed from all the scores: the maximum score over all the k-mers in the sequence (in this case MAX = 7.7841), and the average score over the top 3 highest-scoring k-mers in the sequence—highlighted values (in this case TOP3AVG $= (7.7841 + 5.6353 + 6.7985)/3 = 6.7392$).

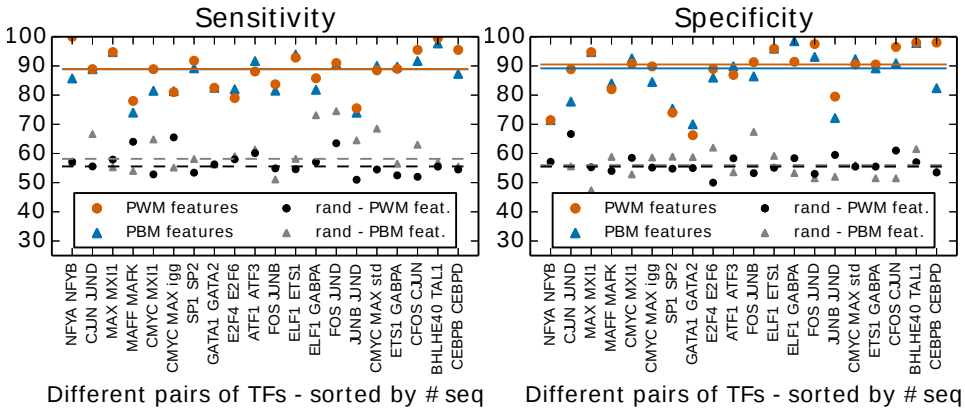


Figure S2. COUGER classification performance (sensitivity & specificity) for 20 pairs of TFs, with PBM and PWM features. The values correspond to the best result from all three classifiers ($\mathbf{SVM_{lin}}$, $\mathbf{SVM_{rbf}}$ and $\mathbf{RF_{pi}}$) and all four sets of features derived by a FS procedure. The horizontal lines represent the median values over all pairs of TFs for each type of feature (PBM- or PWM-derived).

(A)

CMYC-MXI1 - MEDIAN (538 seqs)

Set	Features	SVM-lin	SVM-rbf	RF-pi
ALL	858	88.8889	89.7196	90.74074
RFFSu100	53	89.7196	88.8889	90.65421
RFFSu10	6	88.785	87.8505	87.8505
NMIFS	5	89.7196	87.037	88.8889
NMIFS	10	88.8889	88.8889	87.963

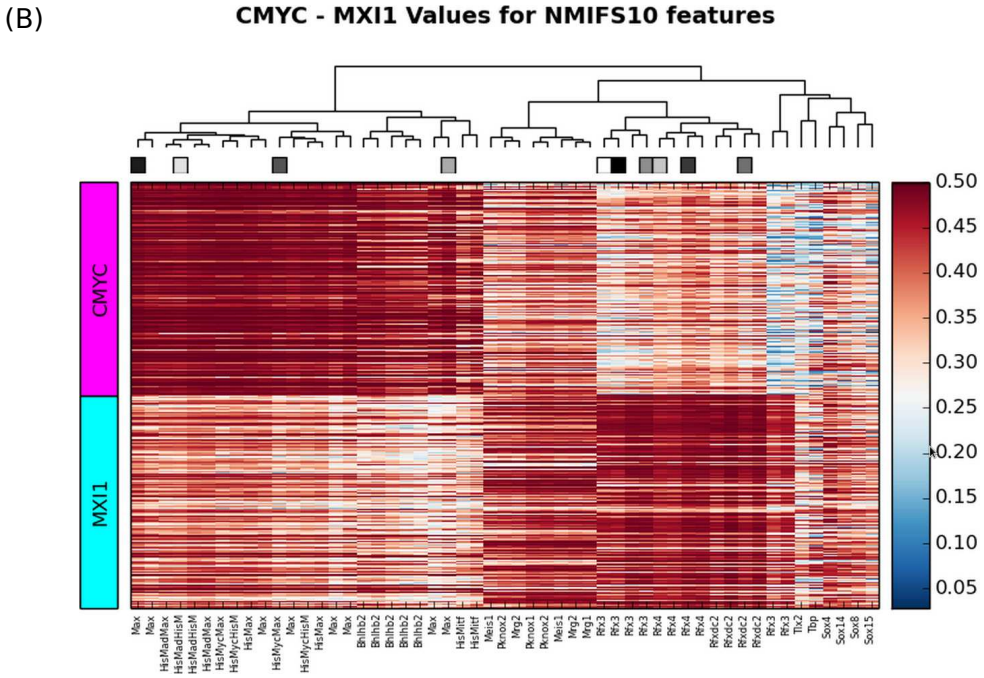
Select algorithm(s): SVM-lin, SVM-rbf, RF-pi

Select feature set(s): RFFS (under 100), RFFS (under 10), NMIFS (10 feat.)

Select run(s): RUN 1, RUN 2, RUN 3, RUN 4, RUN 5

Clear algorithms | Clear sets | Clear runs

(B)



(C)

NMIFS10

ID	NAME	CORR
1	Rfx3_3961_contig8mers_v1_TOP3AVG	0.71
2	Max_3863_contig8mers_v2_MAX	-0.66
3	Rfx4_3761_contig8mers_v1_MAX	0.66
4	HisMycMax_8mers_11111111_TOP3AVG	-0.65
5	Rfxdc2_3516_contig8mers_v2_TOP3AVG	0.64
6	Rfx3_3961_contig8mers_v2_MAX	0.66
7	Max_3864_contig8mers_v2_MAX	-0.68
8	Rfx4_3761_contig8mers_v1_TOP3AVG	0.7
9	HisMadHisMax_8mers_11111111_TOP3AVG	-0.65
10	Rfx3_3961_contig8mers_v2_TOP3AVG	0.7

Figure S3. COUGER output for c-Myc and Mxi1 factors, with PBM-derived features. (A) Median classification accuracies for c-Myc vs. Mxi1 (left), and options for interactive selection (right). (B) Heatmap showing the feature values. Each row represents a DNA sequence in one of the two classes. Each column represents a selected feature from "RF-FS under 100" (i.e., Random Forest feature selection run to select < 100 features). (C) A set of selected features sorted by their score, together with their correlation (i.e., the Pearson correlation coefficient) with the class label. The first class, in this example c-Myc, is considered class "0". The second class, in this example Mxi1, is considered class "1". Thus, a positive correlation for a particular feature suggests that the feature is important for TF2, while a negative correlation suggests that the feature is important for TF1. The name of each feature contains the name of PBM file used to generate that feature, as well as "MAX" or "TOP3AVG", which specify whether the feature represents the score of the best site in each sequence or an average over the top three sites, respectively.

(A)

CMYC-MXI1 - MEDIAN (538 seqs)

Set	Features	SVM-lin	SVM-rbf	RF-pi
ALL	824	83.1776	82.4074	84.1121
RFFSu100	51	84.2593	83.3333	85.0467
RFFSu10	6	84.2593	83.3333	79.6296
NMIFS	5	84.1121	84.2593	84.1121
NMIFS	10	85.1852	86.9159	83.3333

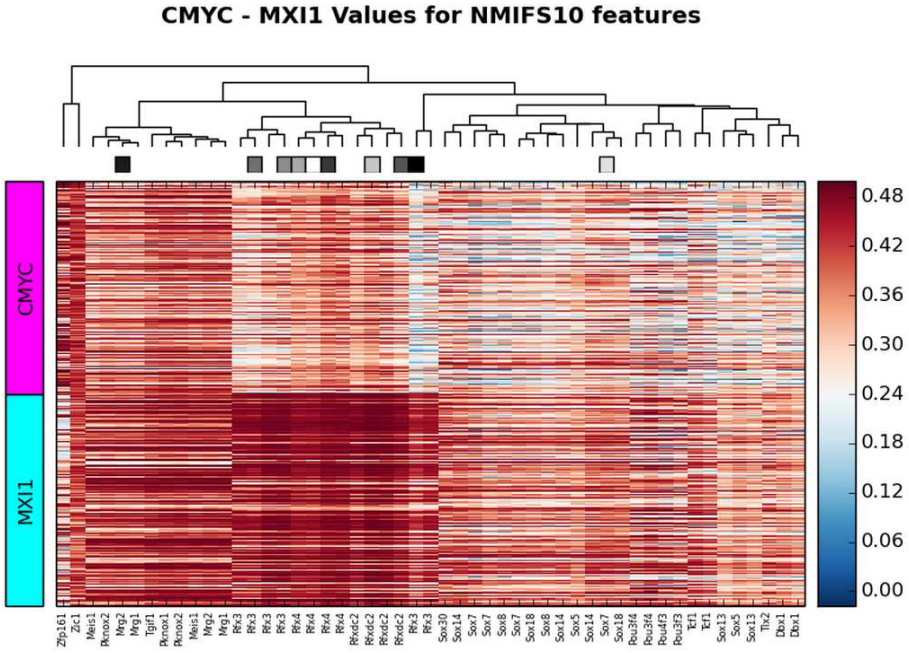
Select algorithm(s): SVM-lin, SVM-rbf, RF-pi

Select feature set(s): RFFS (under 100), RFFS (under 10), NMIFS (10 feat.)

Select run(s): RUN 1, RUN 2, RUN 3, RUN 4, RUN 5

Clear algorithms | Clear sets | Clear runs

(B)



(C)

NMIFS10

ID	NAME	CORR
1	Rfx3_4971_contig8mers_TOP3AVG	0.72
2	Mrg2_2302.1_contig8mers_TOP3AVG	0.32
3	Rfx4_3761_contig8mers_v1_MAX	0.66
4	Rfxdc2_3516_contig8mers_v2_TOP3AVG	0.64
5	Rfx3_3961_contig8mers_v1_TOP3AVG	0.71
6	Rfx3_3961_contig8mers_v2_MAX	0.66
7	Rfx4_3761_contig8mers_v1_TOP3AVG	0.7
8	Rfxdc2_3516_contig8mers_v1_MAX	0.6
9	Sox7_3460_contig8mers_v1_MAX	0.36
10	Rfx4_3761_contig8mers_v2_TOP3AVG	0.69

Figure S4. COUGER output for c-Myc and Mxi1 factors, with PBM-derived features. This figure is simmilar to Fig. **??**, the difference being that in this case the PBM data corresponding to 6 TFs was not considered in classification. The eliminated factors were: c-Myc, Max, Mxi1, Bhlhb2, HisMitf and Tcfe2a.

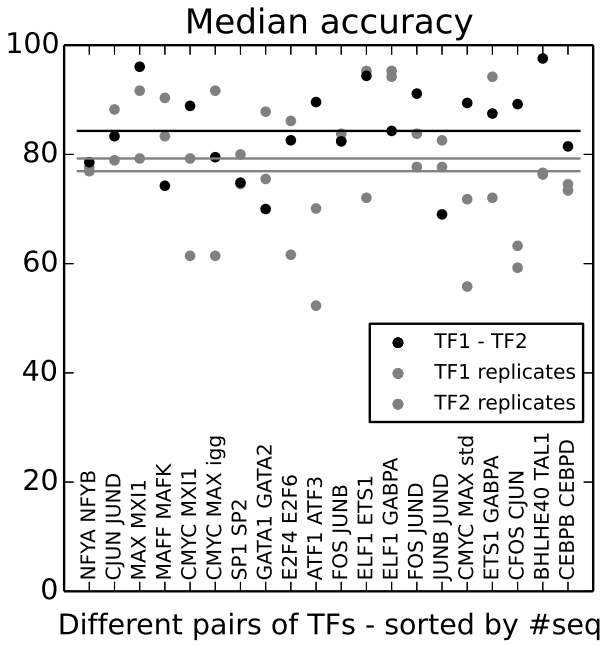


Figure S5. COUGER classification accuracy corresponding to features derived from PBM data from UniPROBE, for 20 pairs of TFs and their pairs of replicates. The values correspond to the median result from all three classifiers ($\mathbf{SVM_{lin}}$, $\mathbf{SVM_{rbf}}$ and $\mathbf{RF_{pi}}$) and all five sets of features (all, under 100 and under 10 features selected by RF-FS, and first 5 and 10 features selected by NMIFS).

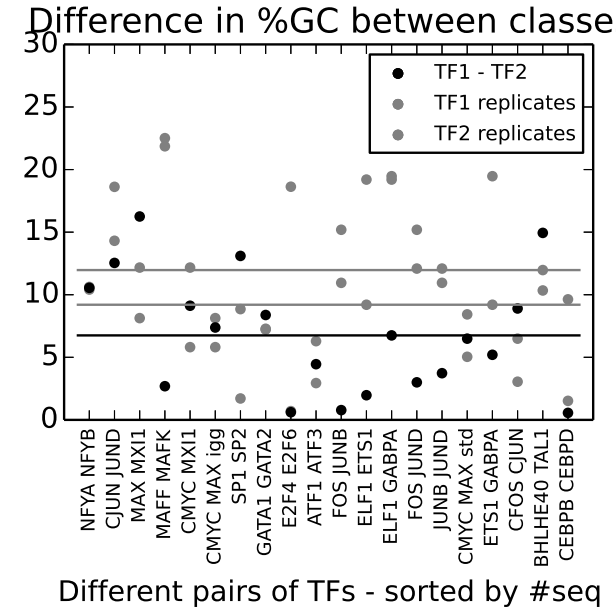


Figure S6. Difference in GC content for all 20 pairs of TFs and their pairs of replicates.

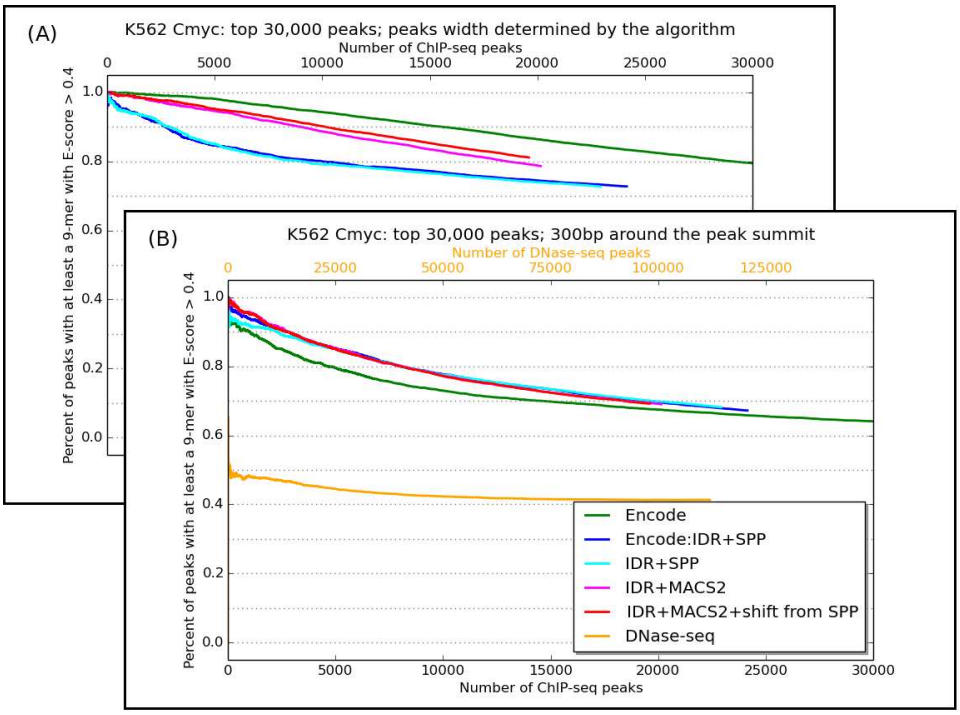


Figure S7. Peak caller comparison for c-Myc dataset: the percent of peaks with at least one high confidence binding site, for (A) peak width determined by the algorithm, and (B) 300bp around the peak summit.

Table S1. Classification accuracies for 20 pairs of paralogous TFs, for PWM-derived features (from UniPROBE, HT-SELEX and JASPAR CORE vertebrata).

			$\mathbf{SVM_{lin}}$					$\mathbf{SVM_{rbf}}$					$\mathbf{RF_{pi}}$					Median
TF1	TF2	#seqs	ALL	FSu100	FSu10	NMIFS5	NMIFS10	ALL	FSu100	FSu10	NMIFS5	NMIFS10	ALL	FSu100	FSu10	NMIFS5	NMIFS10	accuracy
NFYA	NFYB	70	92.86	85.71	64.29	71.43	78.57	85.71	92.86	71.43	78.57	78.57	85.71	85.71	71.43	78.57	78.57	78.57
CJUN	JUND	90	88.89	77.78	83.33	72.22	83.33	88.89	88.89	88.89	72.22	88.89	94.44	94.44	88.89	77.78	88.89	88.89
MAX	MXI1	382	97.4	97.37	94.81	94.81	96.1	96.05	96.05	93.51	94.81	96.1	96.05	96.05	94.81	94.81	97.37	96.05
MAFF	MAFK	506	80.2	75.25	71.29	74.26	77.23	84.16	79.21	76.24	75.25	76.47	81.37	79.21	72.28	75.25	76.47	76.47
CMYC	MXI1	538	91.67	91.67	87.85	90.65	87.96	90.65	93.52	88.89	89.81	88.79	90.74	91.67	87.96	88.79	90.65	90.65
CMYC	MAX	586	81.2	80.34	78.63	79.66	81.36	82.91	82.05	80.51	82.05	78.63	82.2	82.2	80.51	78.81	80.51	80.51
SP1	SP2	736	80.95	78.23	73.47	77.55	76.19	83.67	80.27	72.79	76.87	77.55	78.91	80.95	77.55	78.23	76.87	77.55
GATA1	GATA2	802	70.63	71.88	66.88	70.63	69.38	71.88	71.43	66.88	68.32	68.75	70.19	69.38	67.08	65.22	65.84	69.38
E2F4	E2F6	1002	85.57	82.5	81.59	82.5	81.09	86.57	81.59	81.5	81.5	82	84	82.09	77.5	80.6	80	81.59
ATF1	ATF3	1682	90.18	89.02	86.61	84.82	85.71	91.39	89.91	86.94	86.01	86.05	91.39	90.48	88.1	84.82	86.01	86.94
FOS	JUNB	1842	86.96	85.6	82.34	82.61	83.97	85.91	86.68	82.34	82.93	84.51	85.6	86.41	82.34	83.2	84.78	84.51
ELF1	ETS1	1962	90.05	91.33	90.31	88.8	90.31	92.37	95.66	94.39	93.64	94.15	93.88	94.9	94.39	93.62	94.13	93.64
ELF1	GABPA	1976	81.06	80.76	78.99	76.46	79.24	85.57	88.61	83.04	78.54	81.52	85.82	88.86	83.29	79.24	82.53	81.52
FOS	JUND	2032	93.25	94.5	92.25	93.75	94	94.25	94.75	93.5	93.75	94.75	93.75	95	92.5	94	94.5	94
JUNB	JUND	2194	74.5	71.75	70.75	72.25	71.75	77.5	77	72.75	73.5	76.75	75.25	76.75	71.75	72.75	76.25	73.5
CMYC	MAX	2598	90.5	90.75	87.75	87.75	88.25	90.5	91.75	89.5	89.25	90.75	90.5	91.25	90.25	89.75	90	90.25
ETS1	GABPA	3516	87.75	86.75	85.25	85.5	85.75	90.5	90.25	89.75	89.25	87.75	90.75	90.5	89.25	89.5	89.25	89.25
CFOS	CJUN	5186	97.25	96.75	95.25	86.5	90.25	97.5	97.25	96.5	87	92	97.25	96.75	96.25	85.75	92.25	96.25
BHLHE40	TAL1	7000	99	99.5	98	99	98.75	99	99	97.75	99.25	99	98.5	99.25	98.5	98.75	98.75	99
CEBPB	CEBPD	7152	95	95.25	93.25	94.5	94.75	95.75	97	94.5	96	95.75	96.5	97	94.75	95.5	96.25	95.5

The pairs of factors are sorted by the number of sequences selected by **COUGER** for both classes (TF1- and TF2- specific).

Table S2. Classification accuracies for 20 pairs of paralogous TFs, for PBM-derived features (from UniPROBE).

			$\mathbf{SVM_{lin}}$					$\mathbf{SVM_{rbf}}$					$\mathbf{RF_{pi}}$					Median
TF1	TF2	#seqs	ALL	FSu100	FSu10	NMIFS5	NMIFS10	ALL	FSu100	FSu10	NMIFS5	NMIFS10	ALL	FSu100	FSu10	NMIFS5	NMIFS10	accuracy
NFYA	NFYB	70	85.71	78.57	78.57	78.57	78.57	85.71	85.71	64.29	78.57	78.57	85.71	78.57	78.57	78.57	78.57	78.57
CJUN	JUND	90	83.33	83.33	77.78	83.33	83.33	77.78	88.89	77.78	83.33	83.33	88.89	94.44	83.33	77.78	83.33	83.33
MAX	MXI1	382	97.37	96.1	94.74	96.05	96.05	94.81	96.05	93.42	94.81	96.05	96.1	94.81	94.74	94.74	96.05	96.05
MAFF	MAFK	506	76.24	80.2	71.29	70.3	70.3	77.23	79.21	71.29	74.26	72.28	76.24	78.22	66.34	75.25	73.27	74.26
CMYC	MXI1	538	88.89	89.72	88.79	89.72	88.89	89.72	88.89	87.85	87.04	88.89	90.74	90.65	87.85	88.89	87.96	88.89
CMYC	MAX	586	80.34	78.63	75.21	76.92	78.63	78.63	79.49	79.49	79.49	82.91	80.51	82.91	79.49	82.05	81.36	79.49
SP1	SP2	736	77.55	74.83	73.47	74.15	74.15	77.55	78.91	73.47	73.47	74.83	82.31	80.95	74.32	76.19	80.41	74.83
GATA1	GATA2	802	72.67	73.29	68.94	70.19	70.81	73.13	72.5	68.32	68.94	72.05	69.57	70	65	65.63	69.57	70
E2F4	E2F6	1002	86	83.58	81.5	82	82.59	84.58	83.58	81	81.5	83	84.58	83.08	80.6	82	81.5	82.59
ATF1	ATF3	1682	89.88	89.58	86.31	86.9	87.2	88.69	90.77	87.83	86.61	88.1	91.67	89.88	89.58	90.18	89.58	89.58
FOS	JUNB	1842	80.76	81.25	75.61	80.49	81.84	83.2	82.38	80.22	80.43	82.66	83.47	82.66	83.97	82.93	83.7	82.38
ELF1	ETS1	1962	93.62	94.4	93.88	93.38	94.13	94.13	94.15	94.39	93.64	94.39	94.64	94.64	94.9	94.4	94.9	94.39
ELF1	GABPA	1976	82.03	83.04	82.03	81.01	81.77	81.52	86.33	82.28	84.3	84.81	86.58	87.85	86.58	87.34	88.35	84.3
FOS	JUND	2032	91.13	91.4	90.15	89.9	90.64	91.13	91.89	91.38	89.9	90.89	90.89	92.38	91.87	89.9	91.87	91.13
JUNB	JUND	2194	68.26	67.35	63.33	65.38	66.06	69.02	69.93	69.25	67.65	70.39	71.53	73.35	70.55	66.97	70.55	69.02
CMYC	MAX	2598	89.42	89.79	89.23	84.59	88.27	89.21	90.38	90.17	84.78	89.04	90.75	90.77	89.79	84.81	90.19	89.42
ETS1	GABPA	3516	86.93	85.23	83.36	83.93	84.07	87.77	88.19	84.07	88.9	88.64	89.19	88.64	83.5	87.48	88.92	87.48
CFOS	CJUN	5186	90.45	89.87	82.93	83.22	86.32	91.13	91.71	86.11	85.25	89.21	91.51	91.42	87.28	86.21	90.26	89.21
BHLHE40	TAL1	7000	97.79	98	97.57	94.64	96.07	97.64	98.21	97.71	95	96	98	97.93	97.5	94.79	96.21	97.57
CEBPB	CEBPD	7152	83.02	82.1	78.95	78.55	80.07	84.21	83.43	79.94	79.94	80.64	84.21	84.35	79.51	81.47	82.25	81.47

The pairs of factors are sorted by the number of sequences selected by **COUGER** for both classes (TF1- and TF2- specific).

Table S3. Classification accuracies for 20 pairs of paralogous TFs in the case of randomized class labels, for PWM-derived features (from UniPROBE, HT-SELEX and JASPAR CORE vertebrata).

			$\mathbf{SVM_{lin}}$					$\mathbf{SVM_{rbf}}$					$\mathbf{RF_{pi}}$					Median
TF1	TF2	#seqs	ALL	FSu100	FSu10	NMIFS5	NMIFS10	ALL	FSu100	FSu10	NMIFS5	NMIFS10	ALL	FSu100	FSu10	NMIFS5	NMIFS10	accuracy
NFYA	NFYB	70	57.14	50	50	50	57.14	50	57.14	57.14	42.86	64.29	50	57.14	57.14	42.86	50	35
CJUN	JUND	90	33.33	50	61.11	50	55.56	44.44	50	50	38.89	44.44	55.56	50	50	38.89	50	45
MAX	MXI1	382	48.05	55.26	55.26	53.25	52.63	47.37	52.63	55.26	53.95	53.95	53.95	54.55	50	55.84	51.95	191
MAFF	MAFK	506	53.47	53.47	52.48	51.49	53.47	51.49	53.47	52.48	54.9	55.45	52.48	52.48	51.96	51.49	52.94	253
CMYC	MXI1	538	53.27	54.63	50.93	50	49.07	49.07	51.85	46.73	48.15	46.3	44.86	50.93	50	53.7	52.34	269
CMYC	MAX	586	47.01	47.01	49.57	48.31	50.43	47.86	47.86	51.28	52.14	48.72	49.57	46.15	50.43	49.15	50.43	293
SP1	SP2	736	48.3	49.66	51.02	51.7	50.34	49.66	51.02	48.98	48.98	51.02	51.02	51.02	51.02	52.38	47.62	368
GATA1	GATA2	802	48.45	50	51.25	48.13	49.38	54.38	52.17	46.25	48.13	49.69	49.38	51.25	51.25	50	47.5	401
E2F4	E2F6	1002	49	46.77	48	51.24	47.5	47	46.27	48.76	49.75	47.5	49.5	48	46.5	45.5	48	501
ATF1	ATF3	1682	49.7	49.26	49.4	51.49	48.21	48.66	48.51	48.81	53.27	49.7	47.77	49.4	49.11	51.63	50.45	841
FOS	JUNB	1842	50	48.24	46.88	48.78	50.14	48.78	48.24	46.61	49.73	49.18	51.09	50	49.18	47.97	47.28	921
ELF1	ETS1	1962	51.02	49.49	50.38	49.11	51.28	52.16	51.53	50.89	48.72	50.26	48.21	48.98	48.47	51.28	52.3	981
ELF1	GABPA	1976	49.37	48.35	50.89	49.62	50.63	48.86	49.37	50.38	49.11	49.62	49.75	48.86	50.89	53.28	52.41	988
FOS	JUND	2032	49	51.5	48.75	50.5	49	52	52	51	50.5	49.75	52.25	54	51	50.75	49	1016
JUNB	JUND	2194	52	49.5	53.25	49.75	51	50.5	50.25	51.25	50	51	49.75	49.75	50.5	49.25	50	1097
CMYC	MAX	2598	49.5	48.75	50.25	48.75	48.75	48.5	50	48.5	52.25	50.5	49.25	46.75	51.25	49.75	52.25	1299
ETS1	GABPA	3516	50	50.5	49.75	49.25	49.75	49.25	49.5	47	50.25	49.25	48.75	50.25	48.25	48.75	47.75	1758
CFOS	CJUN	5186	51.5	49	48.25	49	48.25	47	49.5	47	48	48.25	50.25	48.75	49	47.75	46.5	2593
BHLHE40	TAL1	7000	50.25	49	47.75	47.75	49	47.75	50.5	47.25	48.25	48	51	49.25	48.5	52.5	50	3500
CEBPB	CEBPD	7152	51.5	51.25	51.5	48	50.25	49.25	50.25	50.25	48	49.25	50	50.5	50.75	49.75	51.5	3576

The pairs of factors are sorted by the number of sequences selected by **COUGER** for both classes (TF1- and TF2- specific).

Table S4. Classification accuracies for 20 pairs of paralogous TFs in the case of randomized class labels, for PBM-derived features.

			$\mathbf{SVM_{lin}}$					$\mathbf{SVM_{rbf}}$					$\mathbf{RF_{pi}}$					Median
TF1	TF2	#seqs	ALL	FSu100	FSu10	NMIFS5	NMIFS10	ALL	FSu100	FSu10	NMIFS5	NMIFS10	ALL	FSu100	FSu10	NMIFS5	NMIFS10	accuracy
NFYA	NFYB	70	35.71	57.14	42.86	50	50	42.86	57.14	50	57.14	50	50	50	50	50	50	50
CJUN	JUND	90	44.44	61.11	44.44	50	44.44	50	50	50	44.44	55.56	44.44	38.89	55.56	44.44	38.89	44.44
MAX	MXI1	382	59.21	51.32	50	49.35	53.25	57.89	53.25	50.65	50.65	53.25	55.26	50	46.75	48.68	52.63	51.32
MAFF	MAFK	506	51.49	46.53	45.54	50.98	48.04	51.49	48.51	49.51	52.48	47.52	49.02	50.5	49.51	53.47	48.51	49.51
CMYC	MXI1	538	49.53	53.7	56.48	55.14	55.14	48.6	51.4	54.21	53.7	55.56	50.47	55.56	50.93	57.94	53.27	53.7
CMYC	MAX	586	51.28	48.72	51.28	49.15	49.15	51.28	52.14	50.43	47.86	49.15	48.31	48.72	43.59	50.43	48.72	49.15
SP1	SP2	736	52.38	49.66	47.62	52.38	50.34	51.02	49.66	52.7	stable51.02	51.7	48.98	46.94	48.98	51.7	47.3	50.34
GATA1	GATA2	802	42.5	46.58	47.21	46.88	45.34	43.13	46.88	47.83	46.58	50	48.13	47.5	49.07	51.55	51.25	47.21
E2F4	E2F6	1002	47	49.5	48	49.75	48.76	47.76	51	50.75	49	47	49.5	51.24	51.5	49	48.5	49
ATF1	ATF3	1682	49.4	47.48	49.4	50.89	48.81	48.07	51.04	49.7	52.68	48.81	50.89	50.89	50.3	52.23	52.08	50.3
FOS	JUNB	1842	48.91	48.1	46.34	50.41	47.97	50	49.05	47.83	48.91	50.27	49.18	47.97	50.54	51.09	48.78	48.91
ELF1	ETS1	1962	47.7	52.04	49.11	52.55	49.87	50	49.74	46.06	51.02	48.6	51.28	53.18	49.74	51.65	49.74	49.87
ELF1	GABPA	1976	49.11	47.85	51.14	50.13	49.87	47.59	49.87	50.13	49.37	51.65	46.58	51.01	50.89	50.63	52.15	50.13
FOS	JUND	2032	48.25	49.25	49.75	50.25	50.5	50	50.5	50.25	50.5	51.5	50.75	51.5	50	51.25	50	50.25
JUNB	JUND	2194	50.25	48	49.5	49	47.5	49	48.5	49.25	50	47.5	49.25	49.75	50	50.5	50	49.25
CMYC	MAX	2598	48	50	49.25	48	50.5	48.75	49.5	49.25	49.5	49.75	49	49.5	48.75	50	48.5	49.25
ETS1	GABPA	3516	50.5	48.25	49.5	48	50	50.25	50	46.75	48.25	47.5	49.5	49.75	51	50.75	51.25	49.75
CFOS	CJUN	5186	49.75	49.75	48.75	52.25	50.25	48.25	49.75	50.5	52	51	49.25	50.25	49.5	48.75	51.25	49.75
BHLHE40	TAL1	7000	50.5	47.75	47.75	48.5	47.25	47.5	47.25	48.5	48	50.75	49.75	50.25	49.75	50.75	51	48.5
CEBPB	CEBPD	7152	49.75	49.5	50.25	49.75	50.25	51	50.5	50.75	50	50.5	51.75	54.5	50	50.5	51.5	50.5

The pairs of factors are sorted by the number of sequences selected by **COUGER** for both classes (TF1- and TF2- specific).

Table S5. Classification accuracies for 7 pairs of paralogous TFs in the case of 5 different randomizations of class labels, for PBM-derived features.

			$\mathbf{SVM_{lin}}$					$\mathbf{SVM_{rbf}}$					$\mathbf{RF_{pi}}$					Median
TF1	TF2	#seqs	ALL	FSu100	FSu10	NMIFS5	NMIFS10	ALL	FSu100	FSu10	NMIFS5	NMIFS10	ALL	FSu100	FSu10	NMIFS5	NMIFS10	accuracy
CJUN	JUND	90	44.44	61.11	44.44	50	44.44	50	50	50	44.44	55.56	44.44	38.89	55.56	44.44	38.89	44.44
			44.44	55.56	44.44	44.44	50	44.44	50	50	50	50	44.44	55.56	50	44.44	50	50
			61.11	55.56	50	44.44	50	55.56	55.56	55.56	44.44	50	61.11	55.56	44.44	50	50	50
			55.56	50	44.44	44.44	38.89	50	38.89	44.44	50	38.89	50	38.89	44.44	38.89	44.44	44.44
			38.89	44.44	44.44	44.44	38.89	33.33	50	44.44	44.44	44.44	38.89	44.44	44.44	44.44	50	44.44
MAX	MXI1	382	59.21	51.32	50	49.35	53.25	57.89	53.25	50.65	50.65	53.25	55.26	50	46.75	48.68	52.63	51.32
			48.68	44.16	44.16	47.37	46.05	44.16	48.68	48.68	48.68	47.37	46.05	42.11	47.37	44.16	42.86	46.05
			41.56	45.45	46.05	45.45	46.05	42.11	48.05	44.74	44.16	43.42	47.37	42.86	42.86	48.68	50.65	45.45
			46.05	48.05	48.68	53.95	55.26	47.37	49.35	48.68	53.95	55.26	50.65	50.65	47.37	53.25	48.05	49.35
			50	52.63	51.32	47.37	50	50	52.63	47.37	47.37	51.95	53.25	50.65	50	48.68	50.65	50
MAFF	MAFK	506	51.49	46.53	45.54	50.98	48.04	51.49	48.51	49.51	52.48	47.52	49.02	50.5	49.51	53.47	48.51	49.51
			48.04	49.51	46.53	50.5	49.51	47.52	48.51	50.5	50.5	45.54	47.52	49.51	49.51	44.55	45.54	48.51
			44.12	50.5	49.51	50.5	46.53	46.53	47.52	50.5	49.51	47.06	49.51	50.5	50.98	49.51	47.52	49.51
			47.52	51.49	47.52	50.5	47.06	47.52	51.49	46.53	51.49	50.5	47.52	54.46	48.04	50.5	47.06	48.04
			46.53	49.51	51.49	47.52	49.51	53.47	55.45	54.9	53.47	51.49	55.45	55.88	52.48	52.48	48.51	52.48
CMYC	MXI1	538	49.53	53.7	56.48	55.14	55.14	48.6	51.4	54.21	53.7	55.56	50.47	55.56	50.93	57.94	53.27	53.7
			50.93	48.15	52.34	53.7	50.47	50.47	52.78	51.4	53.7	50.93	46.3	52.78	52.34	50.47	51.4	51.4
			49.53	47.22	49.07	48.15	51.85	52.78	48.15	51.4	52.78	50	51.85	47.22	46.73	55.56	51.85	50
			54.63	51.4	51.85	45.37	48.15	49.07	50.47	51.85	46.3	57.41	52.34	52.34	53.7	46.73	51.85	51.85
			51.4	51.85	50.47	56.07	56.48	48.6	53.27	51.85	56.48	49.53	49.07	51.85	51.4	56.07	51.85	51.85
CMYC	MAX	586	51.28	48.72	51.28	49.15	49.15	51.28	52.14	50.43	47.86	49.15	48.31	48.72	43.59	50.43	48.72	49.15
			44.44	43.22	45.76	46.61	47.01	42.37	42.74	43.59	48.72	48.72	47.86	47.86	51.28	48.72	47.86	47.01
			47.86	47.01	46.15	44.44	46.15	45.76	42.74	45.76	46.15	47.46	47.01	48.31	48.72	46.61	44.44	46.15
			51.28	51.28	49.57	47.86	51.28	50.43	48.72	46.15	47.01	47.01	50	46.15	47.86	47.01	46.15	47.86
			49.57	48.72	47.46	46.15	47.01	46.15	48.72	51.28	50.85	47.86	49.57	52.54	50.43	52.14	49.57	49.57
GATA1	GATA2	802	42.5	46.58	47.21	46.88	45.34	43.13	46.88	47.83	46.58	50	48.13	47.5	49.07	51.55	51.25	47.21
			48.13	45	48.75	50.93	51.55	48.13	46.88	48.75	50.31	49.69	50.31	50	51.55	53.13	48.13	49.69
			46.88	49.38	50.31	49.07	49.69	47.83	49.69	46.88	50	52.17	50	49.69	48.75	49.38	47.83	49.38
			48.45	46.58	49.38	45.63	45.63	48.75	49.07	48.45	49.69	50.31	50.31	46.25	51.55	47.83	48.45	48.45
			50.63	51.25	50	53.42	53.75	49.69	49.69	44.72	54.04	52.17	47.21	43.48	48.13	45.96	51.88	50
E2F4	E2F6	1002	47	49.5	48	49.75	48.76	47.76	51	50.75	49	47	49.5	51.24	51.5	49	48.5	49
			51.24	49.75	47.76	49.75	49.5	48.26	51.74	49	50	50.5	49.75	48	51.24	50.5	52	49.75
			47.26	50	48	51.24	51.24	50.5	51	46.5	49.5	50	51	52.5	49.25	47.5	51.24	50
			45.5	50.5	47.5	51.24	48	50.75	50.5	48.76	46.77	50	50	52.5	46.77	47.5	48.5	48.76
			55.22	48.5	47	48.5	47.76	48	48	47.76	47	48.76	48	48	51.24	49	50	48

The pairs of factors are sorted by the number of sequences selected by **COUGER** for both classes (TF1- and TF2- specific).

Table S6. Classification accuracies for pairs of replicates, for PBM-derived features.

TF	$\mathbf{SVM_{lin}}$					$\mathbf{SVM_{rbf}}$					$\mathbf{RF_{pi}}$					Median
	ALL	FSu100	FSu10	NMIFS5	NMIFS10	ALL	FSu100	FSu10	NMIFS5	NMIFS10	ALL	FSu100	FSu10	NMIFS5	NMIFS10	accuracy
CEBPB	74.07	75.93	67.89	71.56	73.39	74.31	76.15	72.48	72.48	72.48	75.23	73.39	67.89	74.07	74.07	73.39
CEBPD	74.74	69.52	65.55	68.27	67.92	82.25	74.53	67.64	70.83	72.29	82.25	79.96	74.95	75.63	77.66	74.53
ELF1	96.57	95.86	93.86	93.36	95.29	96.79	96	94.14	93.71	95.36	96.21	95.64	93.79	94	95.29	95.29
ETS1	72.55	73.04	72.06	70.1	71.57	73.53	73.04	71.57	71.08	74.51	73.53	73.04	68.14	68.14	72.06	72.06
GABPA	95.56	94.74	93.7	92.67	94.22	95.93	95.26	93.63	93.41	94.07	95.11	95.04	93.11	93.19	94.22	94.22
SP1	80	80	73.33	80	80	86.67	85.71	73.33	86.67	80	85.71	85.71	80	86.67	86.67	80
SP2	75.5	75.36	73.43	72.5	74.57	75.64	75.21	73.71	72.43	74.5	75.79	75.93	73	71.93	74.57	74.57
ATF1	66.33	65.98	69.39	70.1	72.16	66.33	68.04	68.37	70.1	71.13	72.16	71.13	70.1	65.98	71.13	70.1
ATF3	52.31	53.08	53.08	51.54	53.08	52.31	52.31	51.54	49.23	50.77	52.31	54.62	52.31	53.08	54.62	52.31
BHLHE40	77.57	78.5	74.77	76.64	77.57	76.64	78.5	74.77	76.64	74.77	76.64	80.37	75.7	74.77	75.7	76.64
CFOS	63.64	62.96	57.41	55.56	58.18	61.11	64.81	61.11	57.41	59.26	61.11	60	56.36	51.85	55.56	59.26
CJUN*	78.93	79.75	78.93	77.89	78.31	79.55	79.13	78.51	78.31	78.51	80.99	80.17	77.48	77.27	79.96	78.93
CJUN	61.22	65.31	59.18	65.31	66	69.39	59.18	61.22	65.31	67.35	63.27	61.22	62	66	63.27	63.27
CMYC*	63.77	59.42	60.87	59.42	62.32	60.87	58.57	63.77	58.57	62.32	62.32	65.22	61.43	59.42	65.22	61.43
CMYC	55.81	53.49	55.81	55.81	58.14	55.81	60.47	55.81	55.81	55.81	62.79	60.47	55.81	58.14	60.47	55.81
E2F4	62.03	61.74	59.01	59.13	61.63	61.05	61.16	61.63	60.17	62.9	63.95	64.24	61.05	61.05	64.24	61.63
E2F6	87.43	86.51	84.44	83.92	86.14	87.62	87.25	84.81	84.63	86.32	87.62	86.51	83.36	83.55	85.58	86.14
GATA1	76.78	75.93	75.36	70.23	76.21	76.64	75.5	75.36	70.66	76.21	77.35	76.64	72.08	68.66	75.21	75.5
GATA2	87.83	89.66	84.48	87.93	88.7	87.83	90.52	85.22	87.07	87.83	89.66	88.79	84.35	86.21	86.21	87.83
JUND	90.2	86.27	82	90	90	92.16	88.24	82.35	90	88.24	90	92	80	82.35	88	88.24
MAFF	92.11	91.67	88.16	89.91	90.79	91.67	92.11	89.04	89.91	90.79	90.35	90.79	89.04	89.47	90.35	90.35
MAFK	85.71	83.33	85.71	83.33	83.33	83.33	83.33	85.71	83.33	83.33	83.33	83.33	83.33	83.33	83.33	83.33
MAX*	76.92	76.92	92.31	92.31	92.31	76.92	84.62	92.31	92.31	91.67	92.31	92.31	84.62	84.62	84.62	91.67
MAX	74.36	74.36	67.95	71.79	71.79	73.08	74.36	69.23	71.79	70.89	74.68	70.89	71.79	69.23	72.15	71.79
MXI1	78.72	78.72	78.72	78.19	78.19	79.79	79.26	78.19	80.85	82.45	80.32	79.26	77.66	79.26	81.91	79.26
NFYA	79.68	74.72	73.81	70.81	73.81	80.59	78.05	76.3	74.94	78.78	81.49	81.72	76.92	73.36	79.19	76.92
NFYB	80.74	78.02	74.71	73.01	76.7	82.1	80.19	75.68	74.51	77.67	78.79	80.16	73.54	74.56	78.02	77.67
TAL1	78.95	78.95	76.32	76.32	78.95	78.95	73.68	76.32	76.32	76.32	78.95	73.68	78.95	73.68	76.32	76.32
EFOS	87.16	87.16	84.46	81.76	83.11	87.84	88.51	83.78	83.78	83.11	85.81	85.81	81.08	82.43	82.43	83.78
EJUNB	83.4	82.66	82.32	81.13	82.15	83.75	83.07	82.49	82.14	83.17	83.66	83.33	82.39	81.4	82.58	82.58
EJUND	81.93	79.05	70.95	75.17	77.2	81.08	79.39	76.01	76.69	77.7	80.91	80.57	73.14	74.83	79.73	77.7

The TFs marked with * correspond to a ChIP-seq data set with IgG control. (The marker is used only to diferentiate identical TFs.)

Table S7. Classification accuracies for 11 pairs of replicates in the case of randomized class labels, for PBM-derived features.

	$\mathbf{SVM_{lin}}$					$\mathbf{SVM_{rbf}}$					$\mathbf{RF_{pi}}$					Median
TF	ALL	FSu100	FSu10	NMIFS5	NMIFS10	ALL	FSu100	FSu10	NMIFS5	NMIFS10	ALL	FSu100	FSu10	NMIFS5	NMIFS10	accuracy
CJUN	48.76	49.38	52.27	49.59	49.17	48.97	48.76	50.83	48.14	48.35	50.21	51.65	51.45	51.86	50.62	49.59
CMYC	55.07	50.72	49.28	52.17	50.72	49.28	52.17	52.17	49.28	50.72	56.52	52.17	53.62	43.48	53.62	52.17
E2F4	49.71	48.55	49.28	49.71	49.71	52.33	48.84	50	48.7	49.71	51.59	49.42	50.43	52.17	48.55	49.71
E2F6	48.61	50.93	48.7	49.54	48.61	48.8	50.28	51.02	50.09	51.76	48.98	50.65	50.65	49.35	51.57	50.09
GATA1	48.86	49.43	50.28	49.29	49.86	48.72	50.85	48.72	49.86	48.86	49.86	48.01	49.86	49.29	50.57	49.43
GATA2	46.09	52.17	45.69	53.04	53.45	50.43	51.72	50.43	50	50.86	48.28	51.72	50.86	49.14	49.14	50.43
JUND	43.14	50.98	49.02	50.98	46	46	47.06	49.02	47.06	48	42	43.14	43.14	47.06	47.06	47.06
MAFF	46.05	49.56	49.12	48.25	50	44.3	46.05	48.25	48.25	48.25	45.61	44.74	48.68	46.93	44.3	48.25
MAFK	66.67	57.14	50	50	50	50	71.43	50	50	50	50	50	33.33	42.86	50	50
MAX	53.85	53.85	53.85	53.85	61.54	46.15	61.54	61.54	53.85	61.54	61.54	61.54	69.23	46.15	46.15	53.85
MXI1	54.26	52.66	51.6	50	52.13	52.66	51.06	50	54.26	52.13	52.13	49.47	51.6	48.4	50.53	51.6

Table S9. Difference in GC content (third column) and median classification accuracy (fourth column) for all pairs of paralogous TFs.

TF1	TF2	%GC difference	Median accuracy
CEBPB	CEBPD	0.56	81.47
E2F4	E2F6	0.61	82.59
FOS	JUNB	0.78	82.38
ELF1	ETS1	1.97	94.39
MAFF	MAFK	2.69	74.26
FOS	JUND	3	91.13
JUNB	JUND	3.73	69.02
ATF1	ATF3	4.45	89.58
ETS1	GABPA	5.2	87.48
CMYC	MAX	6.49	89.42
ELF1	GABPA	6.75	84.3
CMYC	MAX*	7.38	79.49
GATA1	GATA2	8.38	70
CFOS	CJUN	8.9	89.21
CMYC	MXI1	9.12	88.89
NFYA	NFYB	10.55	78.57
CJUN	JUND	12.54	83.33
SP1	SP2	13.1	74.83
BHLHE40	TAL1	14.94	97.57
MAX	MXI1	16.25	96.05
Pearson correlation coefficient between median accuracy and %GC difference			**0.217**

The pairs of factors are sorted by %GC difference, which was computed as the absolute value of the difference between TF1's GC content and TF2's GC content. * indicates IgG control for both factors and is used only to diferentiate identical pairs of TFs.

Table S10. Difference in GC content (second column) and median classification accuracy (third column) for all pairs of TF replicates.

TF	%GC difference	Median accuracy
E2F4	0.7	61.63
CEBPB	1.52	73.39
SP1	1.71	80
ATF3	2.94	52.31
CFOS	3.05	59.26
CMYC	5.04	55.81
CMYC*	5.81	61.43
ATF1	6.29	70.1
CJUN	6.49	63.27
GATA2	7.22	87.83
GATA1	7.28	75.5
MAX*	8.13	91.67
MAX	8.43	71.79
ETS1	9.2	72.06
CEBPD	9.63	74.53
BHLHE40	10.34	76.64
NFYB	10.41	77.67
NFYA	10.62	76.92
EJUNB	10.95	82.58
TAL1	11.97	76.32
EJUND	12.09	77.7
MXI1	12.17	79.26
CJUN*	14.31	78.93
EFOS	15.19	83.78
JUND	18.62	88.24
E2F6	18.63	86.14
GABPA	19.47	94.22
MAFK	21.86	83.33
MAFF	22.51	90.35
Pearson correlation coefficient between median accuracy and %GC difference		**0.707**

The pairs of replicates are sorted by %GC difference, which was computed as the absolute value of the difference between the GC content of one replicate and the GC content of the other replicate. * indicates IgG control and is used only to diferentiate identical TFs.

Table S11. Ratio of IDR scores for STAT1 and STAT2.

TF	Treatment	N_p/N_t	N_1/N_2
STAT1	IFNa30	3.22	1.82
STAT1	IFNa6h	2.73	0.65
STAT2	IFNa30	2.94	8.65
STAT2	IFNa6h	1.87	2.07

N_t is the number of peaks consistent between true replicates, N_p is the number of peaks consistent between pooled pseudo-replicates, N_1 and N_2 are the number of self-consistent peaks for each replicate. The restrictions are that $N_p/N_t < 2$ and $0.5 < N_1/N_2 < 2$.

Table S12. Selected PBM features for Fos and JunD when the comparison is between both factors and between replicates.

Comparison	Selected features	Associated with
Fos-JunD	Atf1_3026_contig8mers_v1_MAX	JunD
	Gata6_3769_contig8mers_v1_TOP3AVG	JunD
	Gmeb1_1745_contig8mers_v1_MAX	JunD
	Hbp1_2241_contig8mers_v1_MAX	Fos
	Jundm2_0911_contig8mers_v2_TOP3AVG	Fos
	Lhx6_3432.1_contig8mers_MAX	Fos
	Meis1_2335.1_contig8mers_TOP3AVG	JunD
	Pbx1_3203.1_contig8mers_MAX	JunD
	Sfpi1_8mers_TOP3AVG	JunD
	Zfp691_0895_contig8mers_v1_TOP3AVG	Fos
Fos-Fos	Dbx1_3486.1_contig8mers_TOP3AVG	Rep1
	Gata3_4971_contig8mers_TOP3AVG	Rep1
	Gata5_3768_contig8mers_v2_TOP3AVG	Rep1
	Gata6_3769_contig8mers_v1_TOP3AVG	Rep1
	HisMyc_8mers_11111111_TOP3AVG	Rep1
	Jundm2_0911_contig8mers_v2_MAX	Rep2
	Obox5_2284.1_contig8mers_TOP3AVG	Rep1
	Six6_2267.5_contig8mers_TOP3AVG	Rep1
	Six6_2267_contig8mers_v2_TOP3AVG	Rep1
	Tbp_pr781_contig8mers_v1_TOP3AVG	Rep1
JunD-JunD	Bcl6b_0961_contig8mers_v1_TOP3AVG	Rep2
	E2F3_3752_contig8mers_v2_TOP3AVG	Rep2
	Gata6_3769_contig8mers_v2_MAX	Rep1
	HisMyc_8mers_11111111_TOP3AVG	Rep1
	Hoxd13_2356.1_contig8mers_TOP3AVG	Rep1
	Irx6_2623.2_contig8mers_TOP3AVG	Rep1
	Sox7_3460_contig8mers_v2_TOP3AVG	Rep1
	Sp4_1011_contig8mers_v1_TOP3AVG	Rep2
	Tcfap2c_2912_contig8mers_v1_MAX	Rep2
	Zic2_2895_contig8mers_v2_TOP3AVG	Rep2

The subsets of factors are in alphabetical order and represent the NMIFS10 sets – the first 10 features selected by NMIFS (normalized mutual information feature selection).

Table S13. Difference in performance for Fos and JunD with PBM features, when the number of TF-specific sequences is reduced.

	Number of sequences in each class			
	1016	800	500	300
Median Accuracy	91.13	91.56	94.00	95.83
Median Precision	90.56	91.75	92.00	90.99
Median Sensitivity	89.16	89.38	92.00	93.33
Median Specificity	90.40	91.88	92.00	91.67

The values correspond to the median result from all three classifiers ($\mathbf{SVM_{lin}}$, $\mathbf{SVM_{rbf}}$ and $\mathbf{RF_{pi}}$) and all five considered sets of features (ALL, FSu100, FSu10, NMIFS5, NMIFS10).

Table S14. Different sets of putative cofactors for Fos and JunD with PBM features, when the number of TF-specific sequences is reduced.

# seqs	Putative cofactors	Associated with
1016	Atf1_3026_contig8mers_v1_MAX	JunD
	Gata6_3769_contig8mers_v1_TOP3AVG	JunD
	Gmeb1_1745_contig8mers_v1_MAX	JunD
	Hbp1_2241_contig8mers_v1_MAX	Fos
	JunDm2_0911_contig8mers_v2_TOP3AVG	Fos
	Lhx6_3432.1_contig8mers_MAX	Fos
	Meis1_2335.1_contig8mers_TOP3AVG	JunD
	Pbx1_3203.1_contig8mers_MAX	JunD
	Sfpi1_8mers_TOP3AVG	JunD
	Zfp691_0895_contig8mers_v1_TOP3AVG	Fos
800	Atf1_3026_contig8mers_v1_MAX	JunD
	Atf1_3026_contig8mers_v1_TOP3AVG	JunD
	Gmeb1_1745_contig8mers_v1_MAX	JunD
	Jun_Fos_8mers_MAX	Fos
	Jun_Fos_8mers_TOP3AVG	Fos
	JunDm2_0911_contig8mers_v2_TOP3AVG	Fos
	Pbx1_3203.1_contig8mers_MAX	JunD
	Pknox1_2364.2_contig8mers_TOP3AVG	JunD
	Zfp691_0895_contig8mers_v1_MAX	Fos
	Zfp691_0895_contig8mers_v2_TOP3AVG	Fos
500	Atf1_3026_contig8mers_v1_MAX	JunD
	Gata5_3768_contig8mers_v2_TOP3AVG	JunD
	Gmeb1_1745_contig8mers_v1_MAX	JunD
	Jun_Fos_8mers_MAX	Fos
	Jun_Fos_8mers_TOP3AVG	Fos
	JunDm2_0911_contig8mers_v2_MAX	Fos
	JunDm2_0911_contig8mers_v2_TOP3AVG	Fos
	Mrg1_2246.2_contig8mers_MAX	JunD
	Zfp691_0895_contig8mers_v1_TOP3AVG	Fos
	Zfp691_0895_contig8mers_v2_TOP3AVG	Fos
300	Atf1_3026_contig8mers_v1_MAX	JunD
	Atf1_3026_contig8mers_v1_TOP3AVG	JunD
	Jun_Fos_8mers_MAX	Fos
	Jun_Fos_8mers_TOP3AVG	Fos
	JunDm2_0911_contig8mers_v2_MAX	Fos
	JunDm2_0911_contig8mers_v2_TOP3AVG	Fos
	Tgif2_3451.1_contig8mers_TOP3AVG	JunD
	Zfp691_0895_contig8mers_v1_MAX	Fos
	Zfp691_0895_contig8mers_v1_TOP3AVG	Fos
	Zfp691_0895_contig8mers_v2_TOP3AVG	Fos

The subsets of putative cofactors are in alphabetical order and represent the NMIFS10 sets – the first 10 features selected by NMIFS (normalized mutual information feature selection).

Table S15. Selected putative cofactors for Fos and JunD when PWM features are used.

Feature set	Putative cofactors	Associated with
PWM features from UniPROBE, HT-SELEX, JASPAR	CREB1_MA0018.2_jaspar_TOP3AVG	JunD
	GATA3_648_selex_MAX	JunD
	Gata4_MA0482.1_jaspar_MAX	JunD
	HLF_MA0043.1_jaspar_MAX	JunD
	Hoxc9_MA0485.1_jaspar_MAX	Fos
	JUN_MA0489.1_jaspar_MAX	Fos
	JUN::FOS_MA0099.2_jaspar_MAX	Fos
	Jundt2_txt_secondary_TOP3AVG	Fos
	Nfe2l2_MA0150.2_jaspar_MAX	Fos
	TAL1::GATA1_MA0140.2_jaspar_MAX	JunD

The subsets of putative cofactors are in alphabetical order and represent the NMIFS10 sets – the first 10 features selected by NMIFS (normalized mutual information feature selection).

REFERENCES

1. Zhang, Y., Liu, T., Meyer, C. a., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.,* **9**(9), R137 [PubMed:18798982] [PubMed Central:PMC2592715] [doi:10.1186/gb-2008-9-9-r137].
2. Kharchenko, P. V., Tolstorukov, M. Y., and Park, P. J. (December, 2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.,* **26**(12), 1351–1359 [PubMed:19029915] [PubMed Central:PMC2597701] [doi:10.1038/nbt.1508].
3. Wilbanks, E. G. and Facciotti, M. T. (2010) Evaluation of algorithm performance in ChIP-seq peak detection.. *PloS one,* **5**(7), e11471 [PubMed:20628599] [PubMed Central:PMC2900203] [doi:10.1371/journal.pone.0011471].
4. Li, Q., Brown, J. B., Huang, H., and Bickel, P. J. (2011) Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.,* **5**(3), 1752–1779 [doi:10.1214/11-AOAS466].
5. ENCODE Project Consortium, Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., and Snyder, M. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature,* **489**(7414), 57–74 [PubMed:22955616] [PubMed Central:PMC3439153] [doi:10.1038/nature11247].
6. Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., and et al. (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.,* **22**(9), 1813–31 [PubMed:22955991] [PubMed Central:PMC3431496] [doi:10.1101/gr.136184.111].