Supporting material for

Parameter-tuning-free identification system of transcriptional regulation motifs in genome DNA sequences based on direct comparison scheme of signal/noise distributions

Ryo Nakaki[1], Jiyoung Kang[1], and Masaru Tateno[2]

[1] Graduate School of Pure Applied Science, University of Tsukuba, 1-1-1 Tennodai, Tsukuba Science City, Ibaraki 305-8577, Japan

[2] Graduate School of Life Science, University of Hyogo, 3-2-1 Kouto, Kamigori, Ako, Hyogo 678-1297, Japan

Corresponding should be addressed to *tateno@sci.u-hyogo.ac.jp*

Table S1. The 65 yeast datasets exploited in this study and the number of DNA fragments in each dataset

To evaluate the accuracy of the six programs (the five existing programs and ours), 65 datasets, which were experimentally extracted from the yeast genome DNA sequences by using the ChIP-on-chip technique in the previous study (1), were exploited. The number of DNA fragments involved in each dataset is distinct.

Table S2. The experimental conditions of the datasets that were used for the identification of mammalian TFBMs in the present study

The accuracy of our system was tested with the use of four mammalian datasets extracted by ChIP-on-chip and ChIP-seq techniques.

Tables S3-S16. Results of the TFBM identification using the six programs

Identification of TFBMs was performed under the following conditions: 1) The reduction of SSRs using RepeatMasker (http://www.repeatmasker.org/) was performed or not. 2) For MDscan, the sequences in either the descending or unsorted order were utilized.

Then, "logo", which represents information contents of the identified PSSMs, was generated by Weblogo (2) for each identified TFBM. Also, the frequency of each identified TFBM (the cluster sizes in our system), the (maximum) correlation coefficient between the corresponding elements of the obtained and reference PSSMs (see the text), the rank, and the identified TFBM are shown in Tables S3-S16.

Table S17. The top rank motifs that were identified using the six programs, with respect to the four mammalian datasets.

Figure S1. Determination of R value

For the determination of the most suitable *R* value (i.e., the ratio of "noise"; see subsection 2.5), various *R* values were tested.

Figures S2-S7. The cumulative frequencies of "Ranks 1-5" of the 65 datasets, with respect to the six programs under the various conditions

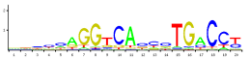
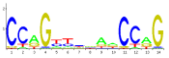
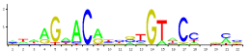
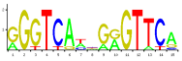
The cumulative frequency $c(r)$ (r represents "Rank", i.e., r=1, …, 5) in equation 13 is plotted with respect to the six programs executed under various conditions.

Table S1. The 65 yeast datasets exploited in this study, and the number of DNA fragments in each dataset.

TF	File name	Number of fragments[a]	TF	File name	Number of fragments
ABF1	ABF1_YPD.fsa	177	RAP1	RAP1_YPD.fsa	108
ACE2	ACE2_YPD.fsa	70	RCS1	RCS1_H2O2Hi.fsa	40
AFT2	AFT2_H2O2Lo.fsa	75	RDS1	RDS1_H2O2Hi.fsa	48
AZF1	AZF1_YPD.fsa	23	REB1	REB1_YPD.fsa	98
BAS1	BAS1_SM.fsa	16	RFX1	RFX1_YPD.fsa	24
CAD1	CAD1_YPD.fsa	28	RLR1	RLR1_YPD.fsa	34
CBF1	CBF1_SM.fsa	194	RPN4	RPN4_H2O2Lo.fsa	69
CIN5	CIN5_H2O2Lo.fsa	117	SFP1	SFP1_SM.fsa	36
DAL82	DAL82_SM.fsa	61	SIG1	SIG1_H2O2Hi.fsa	15
DIG1	DIG1_YPD.fsa	65	SIP4	SIP4_SM.fsa	23
FHL1	FHL1_YPD.fsa	130	SKN7	SKN7_H2O2Lo.fsa	147
FKH1	FKH1_YPD.fsa	103	SNT2	SNT2_YPD.fsa	45
FKH2	FKH2_YPD.fsa	90	SOK2	SOK2_BUT14.fsa	72
GAL4	GAL4_RAFF.fsa	36	SPT23	SPT23_YPD.fsa	47
GAT1	GAT1_RAPA.fsa	48	SPT2	SPT2_YPD.fsa	52
GCN4	GCN4_SM.fsa	142	STB1	STB1_YPD.fsa	22
GLN3	GLN3_RAPA.fsa	78	STB4	STB4_YPD.fsa	27
HAP1	HAP1_YPD.fsa	131	STB5	STB5_YPD.fsa	43
HAP4	HAP4_YPD.fsa	53	STE12	STE12_Alpha.fsa	141
HSF1	HSF1_H2O2Lo.fsa	73	SUM1	SUM1_YPD.fsa	50
INO2	INO2_YPD.fsa	34	SUT1	SUT1_YPD.fsa	75
INO4	INO4_YPD.fsa	31	SWI4	SWI4_YPD.fsa	129
LEU3	LEU3_SM.fsa	31	SWI6	SWI6_YPD.fsa	120
MBP1	MBP1_H2O2Hi.fsa	91	TEC1	TEC1_YPD.fsa	36
MCM1	MCM1_Alpha.fsa	76	THI2	THI2_Thi-.fsa	48
MET4	MET4_SM.fsa	36	TYE7	TYE7_YPD.fsa	65
MSN2	MSN2_H2O2Hi.fsa	73	UME1	UME1_YPD.fsa	35
NDD1	NDD1_YPD.fsa	93	UME6	UME6_YPD.fsa	92
NRG1	NRG1_H2O2Hi.fsa	107	YAP1	YAP1_H2O2Lo.fsa	36
PDR1	PDR1_YPD.fsa	67	YAP7	YAP7_H2O2Hi.fsa	100
PHD1	PHD1_BUT90.fsa	102	YDR026c	YDR026c_YPD.fsa	14
PHO2	PHO2_H2O2Hi.fsa	13	ZAP1	ZAP1_YPD.fsa	17
PHO4	PHO4_Pi-.fsa	23			

[a] The number of target DNA fragments involved in each dataset.

Table S2. The experimental conditions of the datasets used for mammalian TFBM identification in the present study.

TF	Species	Experimental technique[a]	Number of DNA fragments[b]	Reference	Reference TFBM[c] (JASPAR ID[d])
Estrogen receptor (ER)	human	ChIP-on-chip	1910	*Nat Genet*, **38**, 1289-1297 (2006)	 (MA0112.2)
Tcfcp2l1	mouse	ChIP-seq	3842	*Cell*, **133**, 1106-1117 (2008)	 (MA0145.1)
Androgen receptor (AR)	human	ChIP-seq	1160	*EMBO J*, **30**, 3962-3976 (2011)	 (MA0007.1)
Vitamin D receptor (VDR)	human	ChIP-seq	461	*Genome Res*, **20**, 1352-1360 (2010)	 (MA0074.1)

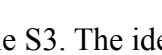
[a] The experimental techniques used for the identification of each TFBM.

[b] The number of DNA fragments of chromosomes 1 and 2.

[c] TFBMs used as the reference.

[d] The JASPAR ID for each reference TFBM.

Table S3. The identification results of our system, without any dataset pre-treatments.

TF	Identified motif	Frequency	Correlation coefficient	Rank
ABF1		211	0.9201	1
ACE2	-	-	-	DA
AFT2		101	0.911	1
AZF1	-	-	-	DA
BAS1		28	0.9529	1
CAD1	-	-	-	DA
CBF1		274	0.98	1
CIN5		161	0.9337	1
DAL82		68	0.9611	2
DIG1		107	0.9803	1
FHL1		172	0.9011	1
FKH1		268	0.9698	1
FKH2		146	0.9778	1
GAL4		49	0.8195	1
GAT1		74	0.8858	4
GCN4		207	0.9481	1
GLN3		130	0.9214	1
HAP1		151	0.8346	1
HAP4		81	0.9561	1
HSF1		100	0.851	1
INO2		55	0.99	1
INO4		52	0.9694	1
LEU3		33	0.8578	1
MBP1		163	0.9325	1
MCM1		220	0.9234	1
MET4		34	0.9233	1
MSN2		88	0.981	1
NDD1	-	-	-	DA
NRG1		129	0.9811	1
PDR1	-	-	-	DA
PHD1		214	0.8417	2
PHO2	-	-	-	DA
PHO4		41	0.9789	1

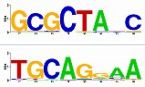
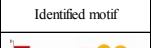
TF	Identified motif	Frequency	Correlation coefficient	Rank
RAP1		143	0.9554	1
RCS1		67	0.9808	1
RDS1		29	0.9801	1
REB1		133	0.9781	1
RFX1		32	0.8853	1
RLR1		123	0.8936	1
RPN4		101	0.9914	1
SFP1		58	0.9762	1
SIG1	-	-	-	DA
SIP4		24	0.8098	2
SKN7		116	0.8382	1
SNT2		35	0.986	1
SOK2		104	0.9367	1
SPT23		108	0.96	2
SPT2	-	-	-	DA
STB1		68	0.8694	1
STB4		37	0.8653	1
STB5		55	0.944	1
STE12		331	0.9748	1
SUM1		39	0.8396	2
SUT1		76	0.9077	2
SWI4		194	0.98	1
SWI6		194	0.8278	1
TEC1		49	0.9156	1
THI2	-	-	-	DA
TYE7		82	0.955	1
UME1	-	-	-	DA
UME6		146	0.9855	1
YAP1		36	0.901	1
YAP7		176	0.9339	1
YDR026c		17	0.996	1
ZAP1		19	0.9518	1

Table S4. The identification results of our system, conducted after the SSR reduction.

TF	Identified motif	Frequency	Correlation coefficient	Rank
ABF1	ATAGTGAT	223	0.9269	1
ACE2	TGCTGGTT	64	0.9038	1
AFT2	GGGTGCA	103	0.9123	1
AZF1	-	-	-	DA
BAS1	TGACTCAT	30	0.9495	1
CAD1	GCTAGTA	30	0.9356	3
CBF1	ATCACGTG	272	0.9772	1
CIN5	TTACATAA	146	0.9281	2
DAL82	CTCTTATC	69	0.9639	2
DIG1	TGTTTCA	80	0.9723	1
FHL1	GTATGGGT	171	0.9032	1
FKH1	TGTTTAC	248	0.971	1
FKH2	GTAAACA	134	0.9751	1
GAL4	-	-	-	DA
GAT1	-	-	-	DA
GCN4	TGACTCAT	220	0.9522	1
GLN3	TCTTATCA	126	0.9511	1
HAP1	ACCGATAA	124	0.824	1
HAP4	TGATTGG	81	0.9561	1
HSF1	TCTAGAA	98	0.8505	1
INO2	CACATGC	57	0.9912	1
INO4	TCACATG	50	0.9814	1
LEU3	GGTTCGC	35	0.8652	1
MBP1	ACGCGT	175	0.9829	1
MCM1	TTTCCAA	213	0.9257	1
MET4	CCACATT	39	0.9275	2
MSN2	CCCCTT	79	0.944	1
NDD1	-	-	-	DA
NRG1	AGGGTCC	118	0.9843	1
PDR1	-	-	-	DA
PHD1	CGCA	227	0.839	2
PHO2	-	-	-	DA
PHO4	CACGTG	35	0.9599	1
RAP1	TGTATGGG	153	0.9584	1
RCS1	GGGTGCA	64	0.979	1
RDS1	CGGCCGA	28	0.963	1
REB1	TTACCCG	134	0.9804	1
RFX1	TGACAAC	30	0.8914	1
RLR1	AAAAATG	114	0.899	3
RPN4	GGTGGCAA	107	0.9914	1
SFP1	CCCATACA	58	0.9747	1
SIG1	-	-	-	DA
SIP4	-	-	-	DA
SKN7	CTGGCCC	114	0.8397	1
SNT2	GCGCTA C	37	0.9839	1
SOK2	TTCTGCA	104	0.9369	1
SPT23	TTATTTC	91	0.9479	5
SPT2	-	-	-	DA
STB1	TTTCGCG	48	0.935	1
STB4	G CGA	28	0.8818	2
STB5	AA CCG	52	0.9409	1
STE12	TGTTTCA	314	0.9769	1
SUM1	TG C C	87	0.9277	1
SUT1	GCGG G	79	0.9042	3
SWI4	TTCGCGTT	149	0.9449	1
SWI6	AAACGCG	182	0.8306	1
TEC1	ACATTCT	50	0.9171	2
THI2	-	-	-	DA
TYE7	CACGTGAT	62	0.9666	1
UME1	-	-	-	DA
UME6	TAGCCGC	139	0.9857	1
YAP1	CTTACAA	71	0.9312	1
YAP7	TTACTAA	151	0.9453	1
YDR026c	CCGGGTA	18	0.9952	1
ZAP1	CCCTAAAG	19	0.9518	1

Table S5. The identification results of MDscan, without any dataset pre-treatments.

TF	Frequency	Correlation coefficient	Rank
ABF1	205	0.8004	1
ACE2	-	-	DA
AFT2	126	0.9676	1
AZF1	-	-	DA
BAS1	34	0.9623	1
CAD1	-	-	DA
CBF1	301	0.9802	1
CIN5	-	-	DA
DAL82	-	-	DA
DIG1	-	-	DA
FHL1	176	0.9904	1
FKH1	139	0.9909	1
FKH2	-	-	DA
GAL4	-	-	DA
GAT1	-	-	DA
GCN4	-	-	DA
GLN3	115	0.9777	1
HAP1	-	-	DA
HAP4	93	0.8413	1
HSF1	-	-	DA
INO2	81	0.8943	1
INO4	51	0.9658	1
LEU3	53	0.9404	1
MBP1	202	0.9873	1
MCM1	160	0.8057	3
MET4	52	0.813	2
MSN2	-	-	DA
NDD1	-	-	DA
NRG1	-	-	DA
PDR1	-	-	DA
PHD1	255	0.915	1
PHO2	-	-	DA
PHO4	47	0.9792	1

TF	Frequency	Correlation coefficient	Rank
RAP1	144	0.9884	1
RCS1	92	0.9514	1
RDS1	-	-	DA
REB1	95	0.9786	1
RFX1	-	-	DA
RLR1	-	-	DA
RPN4	60	0.9982	1
SFP1	53	0.9864	1
SIG1	-	-	DA
SIP4	-	-	DA
SKN7	286	0.8624	2
SNT2	-	-	DA
SOK2	-	-	DA
SPT23	104	0.9248	1
SPT2	-	-	DA
STB1	45	0.9743	1
STB4	-	-	DA
STB5	-	-	DA
STE12	-	-	DA
SUM1	-	-	DA
SUT1	-	-	DA
SWI4	173	0.9498	2
SWI6	253	0.9193	1
TEC1	-	-	DA
THI2	-	-	DA
TYE7	98	0.968	1
UME1	-	-	DA
UME6	84	0.9947	1
YAP1	90	0.8003	1
YAP7	196	0.9828	1
YDR026c	17	0.9898	1
ZAP1	43	0.909	1

Table S6. The identification results of MDscan, conducted after the rearrangement of the target DNA fragments.

TF	Frequency	Correlation coefficient	Rank
ABF1	281	0.9515	1
ACE2	-	-	DA
AFT2	122	0.9837	1
AZF1	140	0.8768	1
BAS1	-	-	DA
CAD1	-	-	DA
CBF1	296	0.9797	1
CIN5	174	0.9924	1
DAL82	114	0.9565	1
DIG1	132	0.9785	1
FHL1	233	0.9643	1
FKH1	148	0.993	1
FKH2	151	0.9234	1
GAL4	45	0.8646	2
GAT1	-	-	DA
GCN4	253	0.965	1
GLN3	100	0.9593	1
HAP1	-	-	DA
HAP4	81	0.9405	5
HSF1	97	0.8429	1
INO2	52	0.9942	2
INO4	52	0.9843	1
LEU3	46	0.9891	1
MBP1	206	0.9831	1
MCM1	167	0.9322	1
MET4	-	-	DA
MSN2	131	0.8591	1
NDD1	-	-	DA
NRG1	173	0.871	1
PDR1	-	-	DA
PHD1	236	0.9116	1
PHO2	-	-	DA
PHO4	38	0.9605	1

TF	Frequency	Correlation coefficient	Rank
RAP1	139	0.9833	1
RCS1	70	0.9763	1
RDS1	45	0.9981	1
REB1	85	0.98	1
RFX1	40	0.916	1
RLR1	-	-	DA
RPN4	65	0.9995	1
SFP1	57	0.9837	1
SIG1	-	-	DA
SIP4	71	0.9551	1
SKN7	251	0.9126	1
SNT2	58	0.9076	1
SOK2	-	-	DA
SPT23	-	-	DA
SPT2	-	-	DA
STB1	60	0.9341	1
STB4	33	0.9032	1
STB5	60	0.9859	2
STE12	225	0.9695	1
SUM1	81	0.9441	1
SUT1	185	0.9005	1
SWI4	277	0.8207	1
SWI6	192	0.8849	1
TEC1	-	-	DA
THI2	-	-	DA
TYE7	87	0.9932	1
UME1	-	-	DA
UME6	80	0.9945	1
YAP1	83	0.8166	5
YAP7	196	0.9967	1
YDR026c	41	0.8427	1
ZAP1	39	0.8543	1

Table S7. The identification results of MDscan, conducted after the SSR reduction.

TF	Frequency	Correlation coefficient	Rank
ABF1	-	-	DA
ACE2	-	-	DA
AFT2	140	0.9717	1
AZF1	-	-	DA
BAS1	45	0.9078	1
CAD1	58	0.978	1
CBF1	300	0.9802	1
CIN5	157	0.9947	1
DAL82	-	-	DA
DIG1	-	-	DA
FHL1	176	0.9904	1
FKH1	-	-	DA
FKH2	119	0.983	1
GAL4	-	-	DA
GAT1	-	-	DA
GCN4	169	0.8675	1
GLN3	-	-	DA
HAP1	227	0.802	5
HAP4	-	-	DA
HSF1	-	-	DA
INO2	74	0.8987	1
INO4	58	0.9559	1
LEU3	42	0.9947	1
MBP1	205	0.9835	1
MCM1	139	0.8708	1
MET4	-	-	DA
MSN2	-	-	DA
NDD1	-	-	DA
NRG1	-	-	DA
PDR1	-	-	DA
PHD1	174	0.8691	2
PHO2	38	0.8287	2
PHO4	39	0.987	1

TF	Frequency	Correlation coefficient	Rank
RAP1	127	0.9585	1
RCS1	-	-	DA
RDS1	48	0.9967	1
REB1	84	0.9798	1
RFX1	53	0.8599	1
RLR1	-	-	DA
RPN4	65	0.9995	1
SFP1	55	0.9882	1
SIG1	-	-	DA
SIP4	-	-	DA
SKN7	290	0.8771	1
SNT2	54	0.9441	1
SOK2	-	-	DA
SPT23	-	-	DA
SPT2	-	-	DA
STB1	56	0.9753	1
STB4	47	0.825	1
STB5	-	-	DA
STE12	249	0.9525	1
SUM1	84	0.9808	1
SUT1	122	0.8211	2
SWI4	-	-	DA
SWI6	156	0.9386	1
TEC1	-	-	DA
THI2	-	-	DA
TYE7	61	0.9903	1
UME1	-	-	DA
UME6	81	0.9952	1
YAP1	-	-	DA
YAP7	145	0.8639	1
YDR026c	24	0.955	5
ZAP1	-	-	DA

Table S8. The identification results of MDscan, conducted after the SSR reduction and the rearrangement of the target DNA fragments.

TF	Frequency	Correlation coefficient	Rank
ABF1	289	0.9512	1
ACE2	-	-	DA
AFT2	126	0.9818	1
AZF1	-	-	DA
BAS1	42	0.92	1
CAD1	65	0.984	3
CBF1	276	0.9737	1
CIN5	170	0.993	1
DAL82	113	0.958	1
DIG1	129	0.9468	2
FHL1	227	0.9666	1
FKH1	145	0.9906	1
FKH2	165	0.9142	1
GAL4	56	0.8669	3
GAT1	-	-	DA
GCN4	254	0.965	1
GLN3	178	0.9092	1
HAP1	-	-	DA
HAP4	81	0.9405	5
HSF1	94	0.8406	1
INO2	46	0.963	1
INO4	46	0.9794	1
LEU3	58	0.9857	1
MBP1	206	0.983	1
MCM1	130	0.9499	1
MET4	-	-	DA
MSN2	138	0.855	1
NDD1	231	0.8074	3
NRG1	176	0.873	1
PDR1	-	-	DA
PHD1	273	0.9336	1
PHO2	-	-	DA
PHO4	39	0.9589	1

TF	Frequency	Correlation coefficient	Rank
RAP1	145	0.9822	1
RCS1	66	0.9774	1
RDS1	55	0.9982	1
REB1	84	0.9798	1
RFX1	44	0.9514	1
RLR1	-	-	DA
RPN4	65	0.9995	1
SFP1	60	0.9828	1
SIG1	-	-	DA
SIP4	66	0.9475	1
SKN7	279	0.9184	1
SNT2	57	0.9085	1
SOK2	-	-	DA
SPT23	-	-	DA
SPT2	-	-	DA
STB1	59	0.9347	1
STB4	29	0.9291	1
STB5	65	0.9826	4
STE12	217	0.9801	1
SUM1	84	0.9456	1
SUT1	185	0.902	3
SWI4	289	0.8142	1
SWI6	193	0.8906	1
TEC1	-	-	DA
THI2	-	-	DA
TYE7	89	0.9936	1
UME1	94	0.8747	2
UME6	80	0.9945	1
YAP1	80	0.885	1
YAP7	195	0.9966	1
YDR026c	-	-	DA
ZAP1	41	0.8591	1

Table S9. The identification results of BioProspector, without any dataset pre-treatments.

TF	Frequency	Correlation coefficient	Rank
ABF1	-	-	DA
ACE2	-	-	DA
AFT2	59	0.9806	1
AZF1	-	-	DA
BAS1	18	0.9505	1
CAD1	-	-	DA
CBF1	155	0.9742	1
CIN5	64	0.865	1
DAL82	-	-	DA
DIG1	34	0.8538	3
FHL1	118	0.9907	1
FKH1	94	0.9897	1
FKH2	60	0.9702	1
GAL4	-	-	DA
GAT1	-	-	DA
GCN4	111	0.9693	1
GLN3	50	0.9576	1
HAP1	77	0.8925	2
HAP4	40	0.9553	1
HSF1	64	0.8253	1
INO2	36	0.9696	1
INO4	34	0.9663	1
LEU3	17	0.9787	1
MBP1	115	0.9717	1
MCM1	54	0.9649	1
MET4	-	-	DA
MSN2	48	0.9699	1
NDD1	-	-	DA
NRG1	84	0.9869	1
PDR1	-	-	DA
PHD1	-	-	DA
PHO2	-	-	DA
PHO4	21	0.9659	1

TF	Frequency	Correlation coefficient	Rank
RAP1	106	0.9938	1
RCS1	40	0.9842	1
RDS1	25	0.9631	1
REB1	95	0.9802	1
RFX1	20	0.9324	1
RLR1	-	-	DA
RPN4	67	0.9994	1
SFP1	40	0.9874	1
SIG1	-	-	DA
SIP4	-	-	DA
SKN7	104	0.901	1
SNT2	23	0.9875	1
SOK2	-	-	DA
SPT23	-	-	DA
SPT2	-	-	DA
STB1	24	0.9033	1
STB4	16	0.8984	2
STB5	28	0.9257	1
STE12	93	0.9806	1
SUM1	44	0.9577	1
SUT1	51	0.8333	3
SWI4	79	0.9357	1
SWI6	105	0.9346	1
TEC1	26	0.926	1
THI2	-	-	DA
TYE7	49	0.9833	1
UME1	-	-	DA
UME6	96	0.9922	1
YAP1	34	0.9691	1
YAP7	91	0.9458	1
YDR026c	13	0.992	1
ZAP1	15	0.8591	2

Table S10. The identification results of BioProspector, conducted after the SSR reduction.

TF	Frequency	Correlation coefficient	Rank
ABF1	-	-	DA
ACE2	-	-	DA
AFT2	55	0.9431	1
AZF1	-	-	DA
BAS1	23	0.9492	1
CAD1	29	0.9902	1
CBF1	142	0.9663	1
CIN5	68	0.9594	1
DAL82	-	-	DA
DIG1	-	-	DA
FHL1	137	0.9903	1
FKH1	90	0.9718	1
FKH2	91	0.9671	1
GAL4	-	-	DA
GAT1	-	-	DA
GCN4	112	0.9698	1
GLN3	48	0.9559	1
HAP1	-	-	DA
HAP4	40	0.9375	1
HSF1	51	0.8243	2
INO2	35	0.9621	1
INO4	29	0.9738	1
LEU3	22	0.9673	1
MBP1	114	0.9711	1
MCM1	54	0.9636	1
MET4	-	-	DA
MSN2	49	0.9813	1
NDD1	-	-	DA
NRG1	89	0.9851	1
PDR1	-	-	DA
PHD1	75	0.8944	2
PHO2	-	-	DA
PHO4	21	0.9567	1

TF	Frequency	Correlation coefficient	Rank
RAP1	107	0.9934	1
RCS1	40	0.9842	1
RDS1	23	0.9522	1
REB1	94	0.9805	1
RFX1	20	0.9263	1
RLR1	-	-	DA
RPN4	65	0.9996	1
SFP1	43	0.9823	1
SIG1	-	-	DA
SIP4	-	-	DA
SKN7	105	0.9134	1
SNT2	22	0.9916	1
SOK2	-	-	DA
SPT23	-	-	DA
SPT2	-	-	DA
STB1	29	0.9203	1
STB4	23	0.8177	1
STB5	26	0.933	1
STE12	87	0.992	1
SUM1	45	0.8792	1
SUT1	58	0.8712	1
SWI4	94	0.9779	1
SWI6	90	0.9161	1
TEC1	-	-	DA
THI2	-	-	DA
TYE7	55	0.9744	1
UME1	-	-	DA
UME6	89	0.9934	1
YAP1	29	0.9824	1
YAP7	82	0.8948	1
YDR026c	14	0.9906	1
ZAP1	-	-	DA

Table S11. The identification results of MEME, without any dataset pre-treatments.

TF	Frequency	Correlation coefficient	Rank
ABF1	-	-	DA
ACE2	-	-	DA
AFT2	52	0.951	1
AZF1	-	-	DA
BAS1	17	0.8644	1
CAD1	-	-	DA
CBF1	168	0.9869	1
CIN5	-	-	DA
DAL82	-	-	DA
DIG1	-	-	DA
FHL1	60	0.9893	1
FKH1	103	0.9879	1
FKH2	-	-	DA
GAL4	-	-	DA
GAT1	2	0.8035	4
GCN4	143	0.9591	1
GLN3	-	-	DA
HAP1	-	-	DA
HAP4	-	-	DA
HSF1	38	0.8417	1
INO2	32	0.945	1
INO4	32	0.9324	1
LEU3	17	0.9772	1
MBP1	92	0.9888	1
MCM1	-	-	DA
MET4	5	0.8428	1
MSN2	2	0.8574	2
NDD1	-	-	DA
NRG1	40	0.977	1
PDR1	3	0.8376	2
PHD1	15	0.9116	3
PHO2	-	-	DA
PHO4	13	0.9531	1

TF	Frequency	Correlation coefficient	Rank
RAP1	72	0.958	1
RCS1	18	0.9628	1
RDS1	16	0.9693	1
REB1	57	0.9752	1
RFX1	14	0.9662	1
RLR1	30	0.8544	2
RPN4	40	0.9971	1
SFP1	20	0.9785	1
SIG1	-	-	DA
SIP4	-	-	DA
SKN7	90	0.9787	1
SNT2	46	0.8717	1
SOK2	28	0.8079	2
SPT23	-	-	DA
SPT2	-	-	DA
STB1	14	0.8967	1
STB4	5	0.856	1
STB5	-	-	DA
STE12	-	-	DA
SUM1	46	0.9621	1
SUT1	31	0.8322	2
SWI4	-	-	DA
SWI6	121	0.9357	1
TEC1	-	-	DA
THI2	-	-	DA
TYE7	37	0.9974	1
UME1	-	-	DA
UME6	47	0.9949	1
YAP1	-	-	DA
YAP7	101	0.9947	1
YDR026c	11	0.9729	1
ZAP1	-	-	DA

Table S12. The identification results of MEME, conducted after the SSR reduction.

TF	Frequency	Correlation coefficient	Rank
ABF1	-	-	DA
ACE2	-	-	DA
AFT2	29	0.9281	1
AZF1	-	-	DA
BAS1	17	0.8644	1
CAD1	-	-	DA
CBF1	163	0.9864	1
CIN5	-	-	DA
DAL82	-	-	DA
DIG1	-	-	DA
FHL1	55	0.9892	1
FKH1	104	0.9849	1
FKH2	-	-	DA
GAL4	-	-	DA
GAT1	2	0.8035	3
GCN4	124	0.9624	1
GLN3	-	-	DA
HAP1	-	-	DA
HAP4	-	-	DA
HSF1	38	0.8417	1
INO2	30	0.9484	1
INO4	30	0.9336	1
LEU3	17	0.9772	1
MBP1	92	0.9909	1
MCM1	-	-	DA
MET4	4	0.8646	1
MSN2	2	0.8939	4
NDD1	-	-	DA
NRG1	40	0.977	1
PDR1	18	0.967	1
PHD1	16	0.8887	5
PHO2	-	-	DA
PHO4	13	0.9531	1

TF	Frequency	Correlation coefficient	Rank
RAP1	72	0.958	1
RCS1	18	0.9628	1
RDS1	16	0.9693	1
REB1	62	0.9755	1
RFX1	13	0.9664	1
RLR1	-	-	DA
RPN4	40	0.9971	1
SFP1	18	0.9757	1
SIG1	-	-	DA
SIP4	-	-	DA
SKN7	98	0.9818	1
SNT2	45	0.8692	1
SOK2	-	-	DA
SPT23	-	-	DA
SPT2	-	-	DA
STB1	14	0.9036	1
STB4	5	0.874	1
STB5	-	-	DA
STE12	2	0.9668	1
SUM1	49	0.9138	1
SUT1	31	0.8181	2
SWI4	24	0.8211	1
SWI6	33	0.8272	1
TEC1	-	-	DA
THI2	-	-	DA
TYE7	38	0.9925	1
UME1	-	-	DA
UME6	47	0.9949	1
YAP1	-	-	DA
YAP7	-	-	DA
YDR026c	11	0.9729	1
ZAP1	-	-	DA

Table S13. The identification results of DME, without any dataset pre-treatments.

TF	Frequency	Correlation coefficient	Rank
ABF1	18	0.888	3
ACE2	23	0.8852	1
AFT2	43	0.9603	1
AZF1	13	0.9365	1
BAS1	9	0.8958	1
CAD1	-	-	DA
CBF1	149	0.9694	1
CIN5	38	0.9575	2
DAL82	-	-	DA
DIG1	14	0.9287	2
FHL1	89	0.9911	1
FKH1	47	0.9713	1
FKH2	40	0.9714	1
GAL4	-	-	DA
GAT1	-	-	DA
GCN4	56	0.9637	1
GLN3	21	0.8598	1
HAP1	-	-	DA
HAP4	26	0.9735	1
HSF1	28	0.8382	1
INO2	23	0.9578	1
INO4	20	0.9819	1
LEU3	22	0.9343	1
MBP1	65	0.9624	1
MCM1	26	0.9279	1
MET4	10	0.8482	3
MSN2	31	0.8937	1
NDD1	-	-	DA
NRG1	63	0.9873	1
PDR1	-	-	DA
PHD1	-	-	DA
PHO2	-	-	DA
PHO4	13	0.9678	1

TF	Frequency	Correlation coefficient	Rank
RAP1	66	0.9831	1
RCS1	26	0.9883	1
RDS1	39	0.8589	1
REB1	89	0.9803	1
RFX1	12	0.9707	1
RLR1	8	0.8477	4
RPN4	67	0.9991	1
SFP1	25	0.9814	1
SIG1	-	-	DA
SIP4	-	-	DA
SKN7	-	-	DA
SNT2	26	0.9496	1
SOK2	-	-	DA
SPT23	-	-	DA
SPT2	-	-	DA
STB1	18	0.9047	1
STB4	-	-	DA
STB5	15	0.8997	1
STE12	47	0.9668	1
SUM1	21	0.9073	1
SUT1	-	-	DA
SWI4	63	0.963	1
SWI6	74	0.8951	1
TEC1	11	0.8398	1
THI2	-	-	DA
TYE7	36	0.9821	1
UME1	10	0.8765	3
UME6	74	0.9923	1
YAP1	17	0.9788	1
YAP7	27	0.8429	1
YDR026c	13	0.9921	1
ZAP1	-	-	DA

Table S14. The identification results of DME, conducted after the SSR reduction.

TF	Frequency	Correlation coefficient	Rank
ABF1	18	0.888	3
ACE2	23	0.8852	1
AFT2	43	0.9603	1
AZF1	12	0.9365	1
BAS1	9	0.8958	1
CAD1	10	0.9773	2
CBF1	159	0.9688	1
CIN5	19	0.9575	1
DAL82	-	-	DA
DIG1	13	0.9071	2
FHL1	91	0.9919	1
FKH1	51	0.9713	1
FKH2	36	0.9714	1
GAL4	-	-	DA
GAT1	8	0.9217	3
GCN4	56	0.9637	1
GLN3	18	0.8598	1
HAP1	34	0.8906	1
HAP4	26	0.9735	1
HSF1	56	0.8382	1
INO2	23	0.9578	1
INO4	20	0.9819	1
LEU3	22	0.9343	1
MBP1	71	0.9614	1
MCM1	26	0.9279	1
MET4	10	0.8482	3
MSN2	31	0.8937	1
NDD1	-	-	DA
NRG1	63	0.9873	1
PDR1	-	-	DA
PHD1	-	-	DA
PHO2	-	-	DA
PHO4	14	0.9663	1

TF	Frequency	Correlation coefficient	Rank
RAP1	68	0.9839	1
RCS1	26	0.9883	1
RDS1	26	0.941	1
REB1	88	0.9801	1
RFX1	12	0.9707	1
RLR1	8	0.8477	4
RPN4	68	0.9991	1
SFP1	25	0.9814	1
SIG1	-	-	DA
SIP4	-	-	DA
SKN7	-	-	DA
SNT2	20	0.991	1
SOK2	-	-	DA
SPT23	8	0.8839	4
SPT2	-	-	DA
STB1	18	0.9047	1
STB4	-	-	DA
STB5	15	0.8997	1
STE12	39	0.9666	1
SUM1	21	0.9073	1
SUT1	-	-	DA
SWI4	61	0.9642	1
SWI6	71	0.9	1
TEC1	11	0.8398	1
THI2	-	-	DA
TYE7	32	0.9801	1
UME1	8	0.8808	5
UME6	74	0.9923	1
YAP1	16	0.979	1
YAP7	27	0.8429	1
YDR026c	13	0.9921	1
ZAP1	-	-	DA

Table S15. The identification results of Weeder, without any dataset pre-treatments.

TF	Frequency	Correlation coefficient	Rank
ABF1	231	0.9575	1
ACE2	-	-	DA
AFT2	83	0.9957	1
AZF1	-	-	DA
BAS1	44	0.9835	1
CAD1	17	0.9417	1
CBF1	344	0.9998	1
CIN5	87	0.9438	1
DAL82	-	-	DA
DIG1	-	-	DA
FHL1	143	0.9893	1
FKH1	635	0.9707	1
FKH2	624	0.9614	1
GAL4	-	-	DA
GAT1	-	-	DA
GCN4	199	0.9922	1
GLN3	277	0.9515	1
HAP1	-	-	DA
HAP4	143	0.9639	1
HSF1	92	0.8422	1
INO2	50	0.9608	1
INO4	30	0.9956	1
LEU3	14	0.9901	1
MBP1	186	0.9845	1
MCM1	158	0.8618	1
MET4	-	-	DA
MSN2	31	0.9555	1
NDD1	58	0.8051	2
NRG1	58	0.9862	1
PDR1	-	-	DA
PHD1	-	-	DA
PHO2	-	-	DA
PHO4	54	0.9911	1

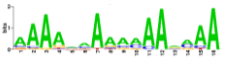
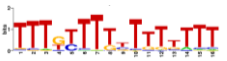
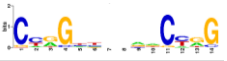
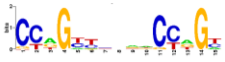
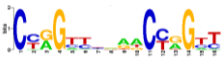
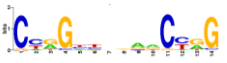
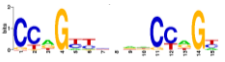
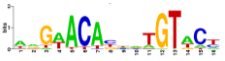
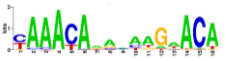
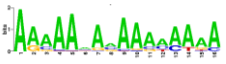
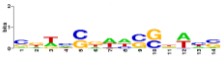
TF	Frequency	Correlation coefficient	Rank
RAP1	55	0.9718	1
RCS1	59	0.9836	1
RDS1	24	0.9925	1
REB1	88	0.9972	1
RFX1	-	-	DA
RLR1	-	-	DA
RPN4	86	0.9959	1
SFP1	35	0.9709	1
SIG1	-	-	DA
SIP4	-	-	DA
SKN7	22	0.8465	4
SNT2	36	0.9762	1
SOK2	142	0.8589	1
SPT23	59	0.8006	2
SPT2	-	-	DA
STB1	46	0.9657	1
STB4	25	0.9224	1
STB5	26	0.9283	1
STE12	352	0.9184	1
SUM1	39	0.9105	2
SUT1	29	0.9151	1
SWI4	83	0.9815	1
SWI6	169	0.9398	1
TEC1	-	-	DA
THI2	-	-	DA
TYE7	108	0.99	1
UME1	-	-	DA
UME6	84	0.9956	1
YAP1	80	0.9475	1
YAP7	201	0.9797	1
YDR026c	29	0.9961	1
ZAP1	-	-	DA

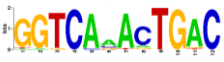
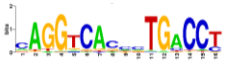
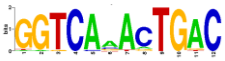
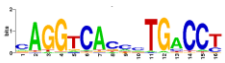
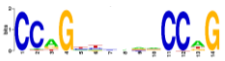
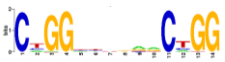
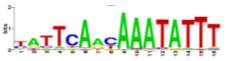
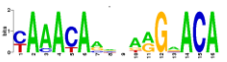
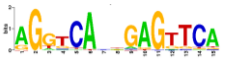
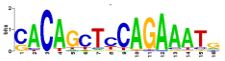
Table S16. The identification results of Weeder, conducted after the SSR reduction.

TF	Frequency	Correlation coefficient	Rank
ABF1	231	0.9575	1
ACE2	-	-	DA
AFT2	83	0.9957	1
AZF1	-	-	DA
BAS1	44	0.9835	1
CAD1	17	0.9417	1
CBF1	344	0.9998	1
CIN5	85	0.9438	1
DAL82	-	-	DA
DIG1	-	-	DA
FHL1	142	0.9895	1
FKH1	603	0.971	1
FKH2	592	0.9618	1
GAL4	-	-	DA
GAT1	-	-	DA
GCN4	197	0.9919	1
GLN3	209	0.9457	1
HAP1	-	-	DA
HAP4	143	0.9639	1
HSF1	78	0.8545	2
INO2	49	0.9627	1
INO4	30	0.9956	1
LEU3	14	0.9901	1
MBP1	183	0.9846	1
MCM1	153	0.8621	1
MET4	-	-	DA
MSN2	31	0.9555	1
NDD1	-	-	DA
NRG1	58	0.9862	1
PDR1	15	0.9497	1
PHD1	-	-	DA
PHO2	-	-	DA
PHO4	54	0.9911	1

TF	Frequency	Correlation coefficient	Rank
RAP1	55	0.9718	1
RCS1	44	0.9601	1
RDS1	22	0.9947	1
REB1	87	0.9972	1
RFX1	20	0.9311	1
RLR1	-	-	DA
RPN4	86	0.9959	1
SFP1	35	0.9955	1
SIG1	-	-	DA
SIP4	-	-	DA
SKN7	-	-	DA
SNT2	36	0.9762	1
SOK2	142	0.8589	1
SPT23	-	-	DA
SPT2	-	-	DA
STB1	46	0.9657	1
STB4	25	0.9224	2
STB5	25	0.9296	1
STE12	322	0.92	1
SUM1	39	0.9105	2
SUT1	29	0.9151	1
SWI4	81	0.9815	1
SWI6	167	0.9398	1
TEC1	31	0.8521	2
THI2	-	-	DA
TYE7	98	0.993	1
UME1	-	-	DA
UME6	84	0.9956	1
YAP1	96	0.9318	1
YAP7	192	0.9796	1
YDR026c	29	0.9961	1
ZAP1	-	-	DA

Table S17. The top rank motifs that were identified using the six programs, with respect to the four mammalian datasets.

TF	SSR reduction	MDscan		MEME		BioProspector	
		Motif	Frequency	Motif	Frequency	Motif	Frequency
hER	×		362		288		277
	○		779		187		667
mTcfcp2l1	×		125		3811		2622
	○		917		2683		2297
hAR	×		328		823		764
	○		228		100		426
hVDR	×		55		110		83
	○		362		3		236

TF	SSR reduction	Weeder		DME		MODIC	
		Motif	Frequency	Motif	Frequency	Motif	Frequency
hER	×		23		16		247
	○		23		16		224
mTcfcp2l1	×		125		8		1322
	○		103		242		1100
hAR	×		91		43		152
	○		12		66		159
hVDR	×		16		11		60
	○		4		5		29

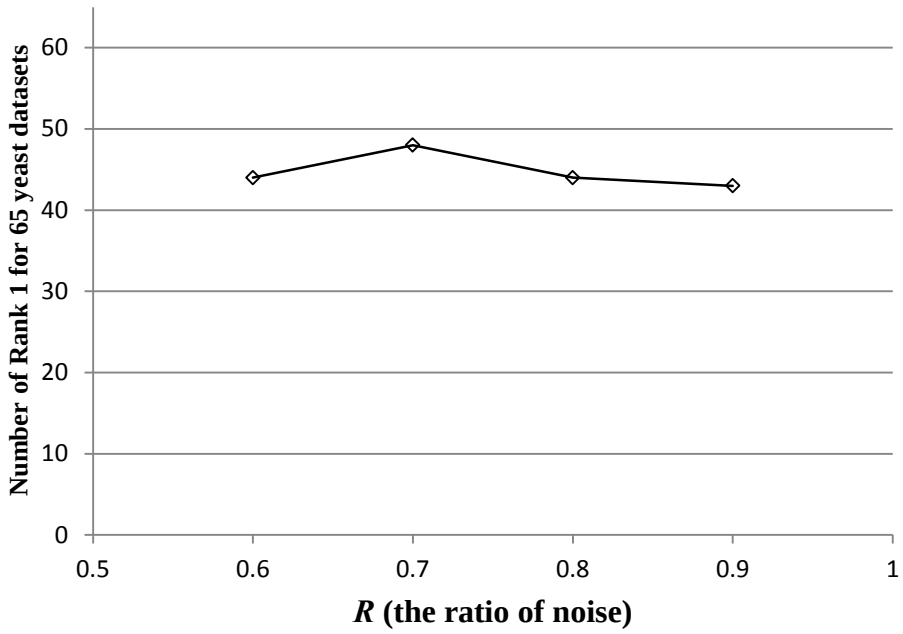


Figure S1. For the determination of the most suitable *R* value (i.e., the ratio of "noise"; see subsection 2.5), 0.6, 0.7, 0.8, and 0.9 were tested. As a result of the analysis, 0.7 was selected such that the number of "Rank 1" was maximized for all 65 yeast datasets.

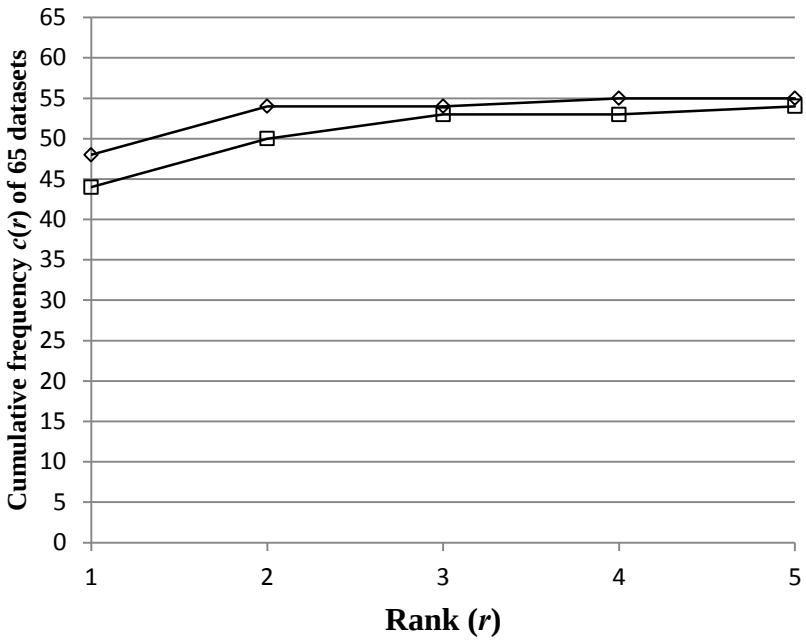


Figure S2. The cumulative frequency $c(r)$ (r represents "Rank", i.e., r=1, …, 5) from equation 13 is plotted with respect to our system; i.e., $c(r)$ of our system, without any dataset pre-treatments (rhombuses), and $c(r)$ of our system conducted after the SSR reduction (squares).

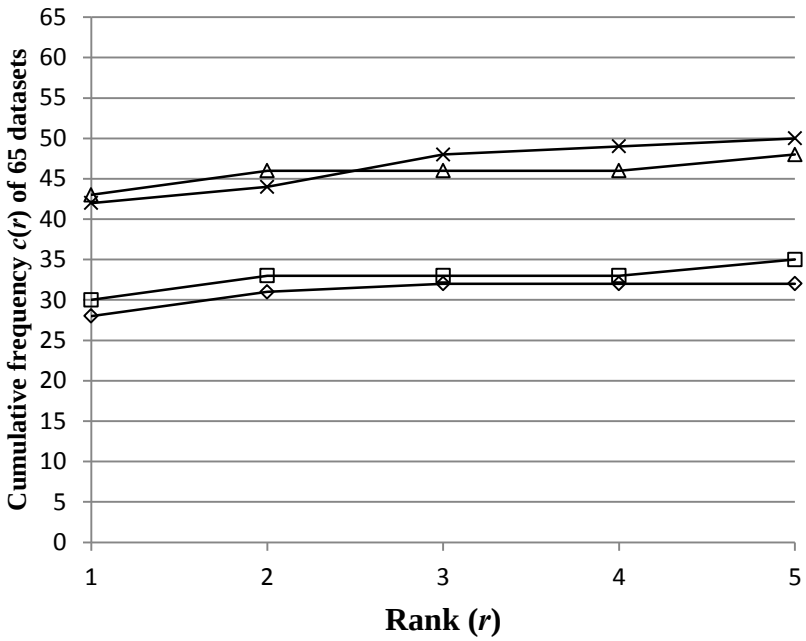


Figure S3. The cumulative frequency $c(r)$ (r represents "Rank", i.e., r=1, …, 5) from equation 13 is plotted with respect to our system; i.e., $c(r)$ of MDscan without any dataset pre-treatments (rhombuses), $c(r)$ of MDscan conducted after the rearrangement of the target DNA fragments (triangles), $c(r)$ of MDscan conducted after SSR reduction (squares), and $c(r)$ of MDscan conducted after the SSR reduction and the rearrangement of the target DNA fragments (crosses).

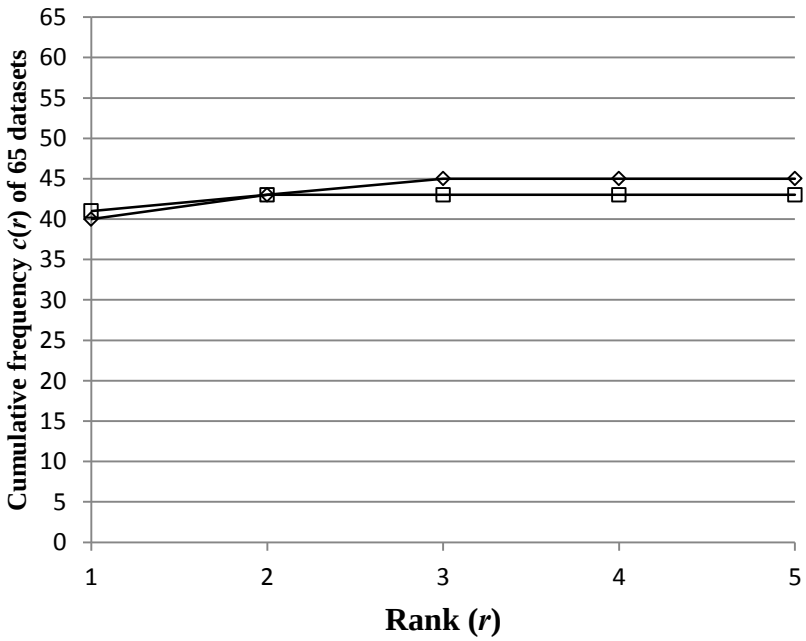


Figure S4. The cumulative frequency $c(r)$ (r represents "Rank", i.e., r=1, …, 5) from equation 13 is plotted with respect to BioProspector; i.e., $c(r)$ of BioProspector without any dataset pre-treatments (rhombuses) and $c(r)$ of BioProspector conducted after the SSR reduction (squares).

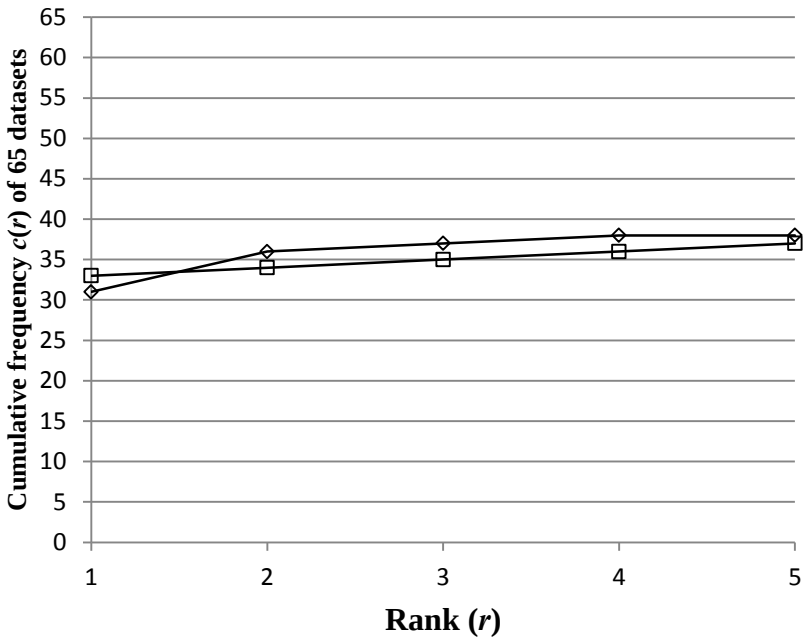


Figure S5. The cumulative frequency $c(r)$ (r represents "Rank", i.e., r=1, …, 5) from equation 13 is plotted with respect to MEME; i.e., $c(r)$ of MEME without any dataset pre-treatments (rhombuses) and $c(r)$ of MEME conducted after the SSR reduction (squares).

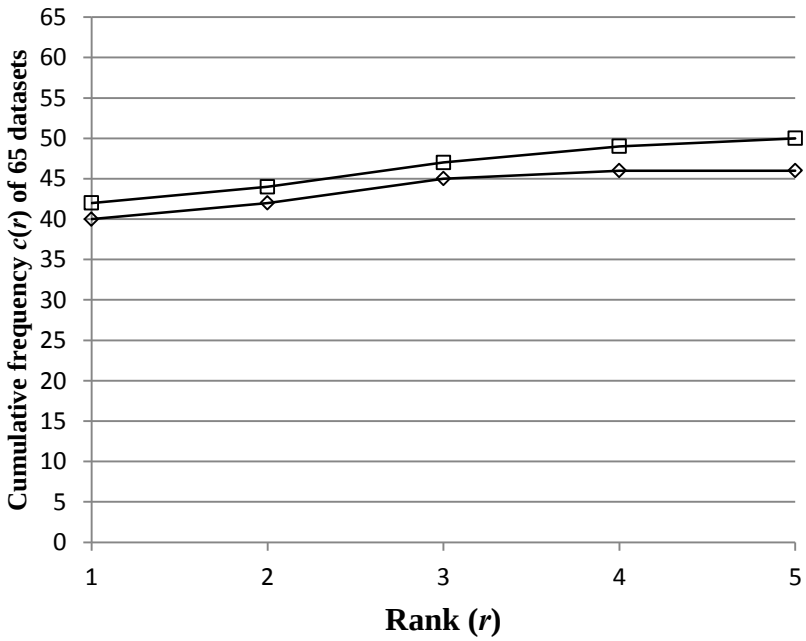


Figure S6. The cumulative frequency $c(r)$ (r represents "Rank", i.e., r=1, …, 5) from equation 13 is plotted with respect to DME; i.e., $c(r)$ of DME without any dataset pre-treatments (rhombuses) and $c(r)$ of DME conducted after the SSR reduction (squares).

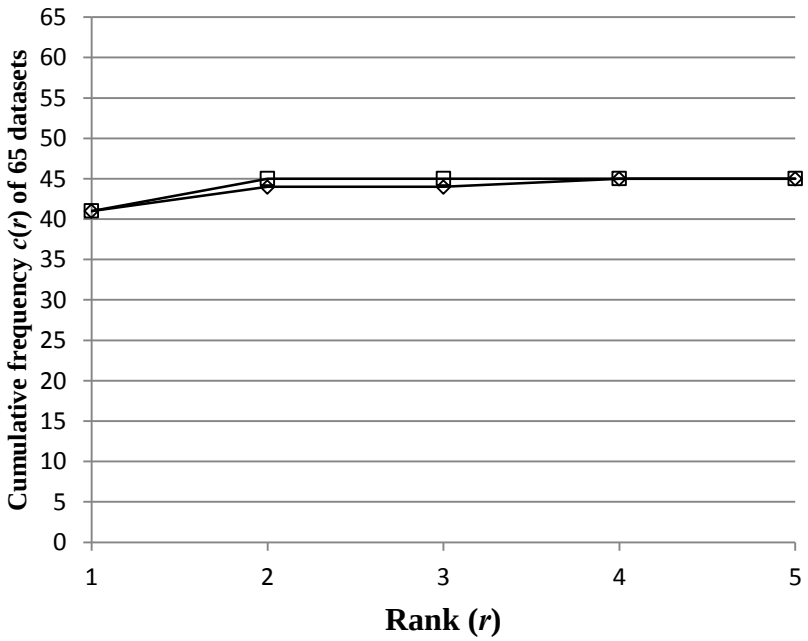


Figure S7. The cumulative frequency $c(r)$ (r represents "Rank", i.e., r=1, …, 5) from equation 13 is plotted with respect to Weeder; i.e., $c(r)$ of Weeder without any dataset pre-treatments (rhombuses) and $c(r)$ of Weeder conducted after the SSR reduction (squares).

References

1. Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99-104.
2. Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) WebLogo: a sequence logo generator. *Genome Res*, **14**, 1188-1190.