Supplementary Information: Discovering regulatory motifs in the Plasmodium genome using comparative genomics

Jie Wu[1,4], Douglas H. Sieglaff[2,3], Joshua Gervin[1] and Xiaohui S. Xie[1,3]

[1]Department of Computer Sciences, University of California, Irvine, CA
[2]Department of Molecular Biology and Biochemistry, University of California, Irvine
[3]Institute of Genomics and Bioinformatics, University of California, Irvine
[4]CODA Genomics, Laguna Hills, CA

1 Generalized hypergeometric distribution

1.1 Basic definition

Consider an urn containing N balls of different sizes, colored with either red or green. Number the balls from i to N. Denote the size of the ith ball by p_i and its color by z_i, with $z_i = 1$ representing red and $z_i = 0$ representing green. We will define a generalized hypergeometric model as a process of randomly picking K balls from the urn without replacement, with the probability of each ball being chosen proportional to the size of the ball. Without loss of generality, we assume p_i is normalized, that is, $\sum_{i=1}^{N} p_i = 1$. A generalized hypergeometric distribution is defined as the probability on the number of the chosen balls that are red.

Let us denote the indices of K chosen balls in one trial by $(X_1, X_2, \cdots, X_K)$, which are random variables with $X_i \in \{1, \cdots, N\}$. The number of the K chosen balls that are red can then be expressed as

$$Y = \sum_{i=1}^{K} z_{X_i} \tag{1}$$

The probability of picking y green balls is then $P(Y = y \mid \mathbf{p}, \mathbf{z}, N, K)$, a distribution parametrized by N, K, $\mathbf{s} = (s_1, \cdots, s_N)$ and $\mathbf{z} = (z_1, \cdots, z_N)$.

The standard hypergeometric distribution can be viewed as a special case of the above distribution, where the sizes of the balls are of equal size, i.e., $p_i = 1/N$ for all i. In this case, the distribution $P(Y)$ can be written down explicitly

$$P(Y = y) = \frac{\binom{M}{y}\binom{N-M}{K-y}}{\binom{N}{K}} \tag{2}$$

where $M = \sum_{i=1}^{N} z_i$ is the total number of red balls.

For other more general forms of $\mathbf{p}$, the distribution $P(Y = y \mid \mathbf{p}, \mathbf{z}, N, K)$ is more complicated and can not be expressed into a simple form. However, the distribution will asymptotically approach the normal distribution when K is large. If we know the mean and variance of the distribution, we will have a good characterization of the probability function. Next we aim to derive the mean and the variance of the distribution.

1.2   Calculation of mean

Since Y is a sum of K random variables, its mean can be expressed as

$$E[Y] = \sum_{i=1}^{K} E[z_{X_i}] \tag{3}$$

where X_i is a random variable representing the index of the ith chosen ball. By definition, the marginal probability of X_i is $P(X_i = k) = p_k$ for all $k = 1, \cdots, N$, and is the same for all i (Note that here we have assumed $\mathbf{p}$ is normalized to be sum 1). Therefore, z_{X_i} is a Bernoulli variable with expectation

$$\phi \equiv E[z_{X_i}] = P(z_{X_i} = 1) = \sum_{i=1}^{N} z_i p_i \tag{4}$$

Accordingly, the mean of Y can be expressed as

$$\mu(N, K, \mathbf{p}, \mathbf{z}) \equiv E[Y] = K\phi = K \sum_{i=1}^{N} z_i p_i \tag{5}$$

1.3   Calculation of variance

The second moment of Y can be expressed as

$$E\left[Y^2\right] = E\left[\sum_{i=1}^{K} z_{X_i}^2 + 2\sum_{i=1}^{K}\sum_{j=i+1}^{K} z_{X_i} z_{X_j}\right] \tag{6}$$

$$= \sum_{i=1}^{K} E\left[z_{X_i}^2\right] + 2\sum_{i=1}^{K}\sum_{j=i+1}^{K} E\left[z_{X_i} z_{X_j}\right] \tag{7}$$

The first part of the above equation is the second moment of z_{X_i}. Since z_{X_i} is a Bernoulli variable, its second moment is equal to its expectation, i.e. ϕ. For the second part, we

notice that

$$
\begin{aligned}
E\left[z_{X_i} z_{X_j}\right] &= P(z_{X_i} = 1, z_{X_j} = 1) & (8) \\
&= \sum_{k=1}^{N} P(z_{X_i} = 1, z_{X_j} = 1 | X_i = k) \; p_k & (9) \\
&= \sum_{k=1}^{N} P(z_{X_j} = 1 | X_i = k) \; z_k \; p_k & (10) \\
&= \sum_{k=1}^{N} \frac{\phi - p_k}{1 - p_k} z_k p_k & (11)
\end{aligned}
$$

Therefore, the second moment of Y is equal to

$$
E[Y^2] = K\phi + K(K-1) \sum_{k=1}^{N} p_i z_k (\phi - p_k)(1 - p_k)^{-1} \tag{12}
$$

Consequently, the variance of Y is

$$
\sigma^2(N, K, \mathbf{p}, \mathbf{z}) \equiv \mathrm{Var}[Y] = K\phi + K(K-1) \sum_{k=1}^{N} (\phi - p_k) p_k (1 - p_k)^{-1} z_k - K^2 \phi^2 \tag{13}
$$

In the special case of the standard hypergeometric distribution ($p_i = 1/N$ for all i), $\phi = M/N$ and consequently the mean is $E[Y] = MK/N$ and the variance can be simplified to $\mathrm{Var}[Y] = MK(1 - M/N)(1 - K/N)/(N - 1)$, both of which can also be derived directly from the standard hypergeometric distribution Eq. (2).

1.4   Simulation studies

To verify the calculation of the mean in Eq. (5) and the variance in Eq. (13), we performed a simulation study. In the study, we took $N = 1000$ and set the first $M = 100$ balls to be red. We did 200 simulation studies with randomly sampled values of $\mathbf{p}$. At each simulation, components of $\mathbf{p}$ are randomly chosen from a uniform distribution ranged between $[0, 1]$. We then sampled $K = 200$ balls according to the generalized hypergeometric process and counted the number of red balls (y). The sampling process was repeated 1000 times, and the empirical mean and variance of y were computed from the sampled data.

The mean and variance for 200 simulation studies are shown in Figure 1. The results demonstrate a good consistency with the theoretical calculations. The figure shows a significant fluctuation in the estimation of variance from sampled data even with 1000 samples, indicating that we will need a much larger number of samples to get a consistent result when using sampling.
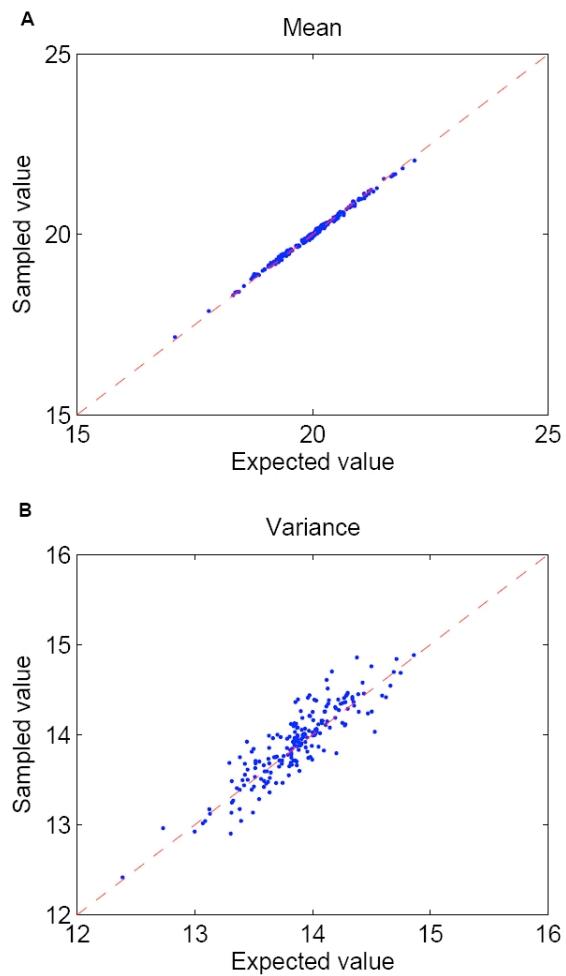
Figure 1: Comparison of mean (A) and variance (B) estimated from sampling to those calculated from theoretical derivations. X-axis shows the sampling results, and Y-axis shows its corresponding theoretical calculations.

2 Application of generalized hypergeometric distribution for motif discovery

The generalized hypergeometric distribution that we have defined above provides a flexible statistical framework for motif discovery using orthologous sequences. Here we describe how to fit the motif discovery scheme into the generalized hypergeometric distribution.

Suppose we are provided with N pairs of orthologous sequences from species A and B. Consider a particular motif m that occurs in n_A sequences in species A, n_B sequences in species B, and in n_{AB} sequences in both A and B. In other words, n_{AB} is the number of sequences containing conserved instances of the motif m, and we want to find out whether n_{AB} is much higher than what is expected by chance.

The MDOS algorithm presented in the main paper defines a background model for calculating the probability of a motif occurring in each of the sequences in species B. The background model considers both the sequence composition and size of each sequence as well as neutral conservation level between orthologous sequence pairs. Based on MDOS, we can calculate the probability vector $\mathbf{p}$ for the motif m to occur in each of the sequences in species B.

Connecting to the generalized hypergeometric distribution, N can be viewed as the total number of balls in an urn, $M = n_A$ is the total number of red balls, $K = n_B$ is the number of balls chosen from the urn, and $y = n_{AB}$ is the number of chosen balls that are red. Vector $\mathbf{p}$ specifies the probability of choosing each ball. The binary vector $\mathbf{z} = (z_1, \cdots, z_N)$ is defined as $z_i = 1$ if the sequence i in species A contains the motif and 0 otherwise. Under this setting, the z-score characterizing the conservation of the motif is defined as

$$\text{zscore} = \frac{n_{AB} - \mu(N, n_B, \mathbf{p}, \mathbf{z})}{\sigma(N, n_B, \mathbf{p}, \mathbf{z})} \tag{14}$$

Note that the above scoring scheme is parametrized on $\mathbf{p}$ and is independent on how $\mathbf{p}$ is chosen. It is possible to extend the MDOS algorithm to consider a more realistic background model for calculating $\mathbf{p}$. The generalized hypergeometric distribution is flexible in incorporating such changes.

3 Application of the MDOS algorithm for discovering motifs in model organisms

We applied the MDOS algorithm to discover motifs in three model organisms – yeast, worm and fly:

Yeast 4,358 orthologous 1Kb promoter sequences between *S. cerevisiae* and *S. bayanus*.

Worm 10,894 orthologous 2Kb promoter sequences between *C. elegans* and *C. briggsae*.

Fly 11,306 orthologous 2Kb promoter sequences between *D. melanogaster* and *D. pseudoobscura*.

The promoter sequences of these organisms were extracted from Elemento et al. (Genome Biology, 2005) and are also available at the supplementary website of this paper.

The distribution of conservation z-scores for 7-mers show significant enrichment with large z-scores in each of the organisms (Figure 2). Note that the bulk of the distribution is centered around 0, indicating that most of the k-mers are indeed neutral. This is in contrast to the z-scores calculated from the standard hypergeometric distribution. Because it does not correct for neutral conservation due to shared ancestry, the bulk of the hypergeometric z-scores is shifted toward the positive side.

The MDOS algorithm calculated the conservation z-scores for all 7, 8 and 9-mers, and expanded significant k-mers to include degenerate codes. The high scoring motifs were further clustered to remove redundancy. The final list of distinct motifs discovered in each of the organisms is shown in the supplementary tables.

Unlike *Plasmodium*, the transcriptional regulation in the genomes of these three model organisms is well studied. In particular, a good number of motifs have been previously described for each of the organisms. Therefore, these organisms provide a good system to test how well the MDOS algorithm can discover real motifs.

We curated a list of known motifs for each of the three organisms from the Transfac database. Overall, we curated 44 motifs for yeast, 10 motifs for worm and 65 motifs for fly.

To test the performance of the MDOS algorithm for identifying real motifs, we focused on examining the properties of 7-mers. For each of the organisms, we first ranked all 7-mers according to their conservation z-scores from high to low. We then systematically identified the 7-mers that match the Transfac motifs using a motif comparing algorithm described previously (Xie et al, 2005), and checked whether the ranks of these 7-mers are biased. A bias toward lower ranks would suggest that the matching 7-mers are enriched with high conservation z-scores. We used Kolmogorov-Smirnov statistics to test such enrichment. For all three organisms, we found a significant enrichment of the matching 7-mers with low ranks (Figure 3). The enrichment p-values for the three organisms are $3.8e^{-14}$,$6.8e^{-15}$, and $2.2e^{-34}$ for yeast, worm and fly respectively, demonstrating that the MDOS algorithm is able to separate functional motifs from neutral ones.
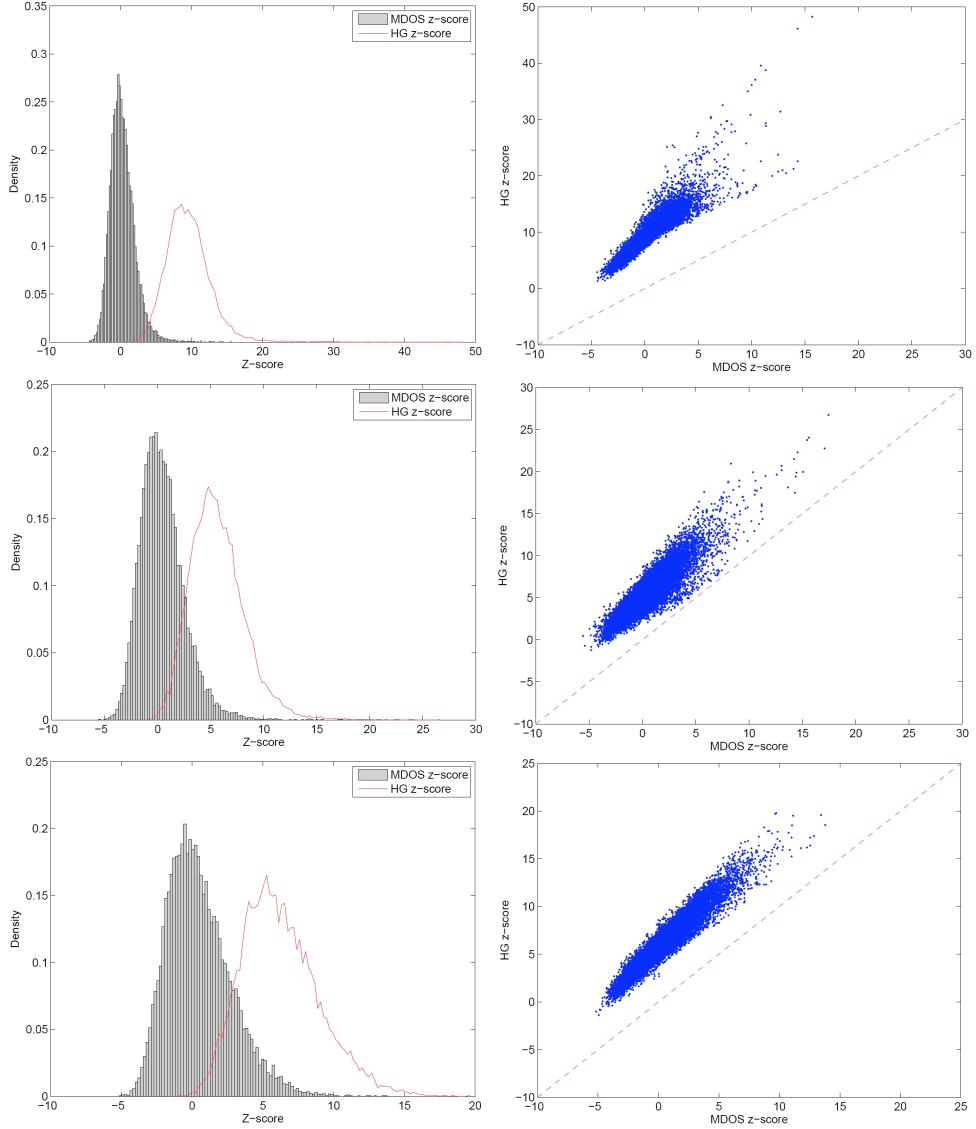
Figure 2: Distribution of conservation z-scores for all 7-mers in *S. cerevesia* (top two panels), *C. elegans* (middle two panels) and *D. Melangaster* (bottom two panels). Left panels show the histograms of z-scores calculated by MDOS and the hypergeometric distribution. Right panels show a comparison of the two scores for each 7-mer. HG z-score represents the hypergeometric z-score.
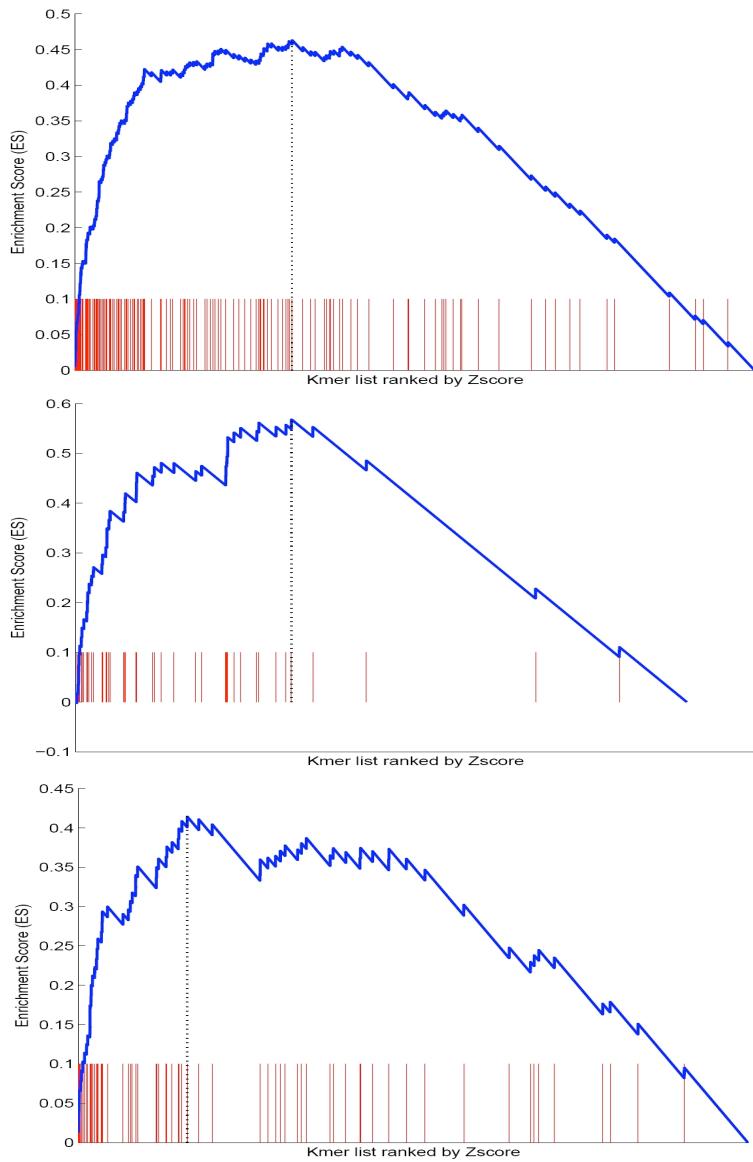
Figure 3: Enrichment of the 7-mers that match to known motifs with high z-scores in three species – yeast (top panel), worm (middle) and fly (bottom). Red vertical line shows the location of the 7-mers matching to known motifs. Blue curves plot the Kolmogorov-Smirnov enrichment scores of these 7-mers. Black line indicates the location of the maximum enrichment score. All three plots show a significant enrichment of the 7-mers matching to known motifs with high z-scores.

8

Table S1 Enrichement of discovered motifs in functional gene sets (21 OPI clusters reported by Young et al, 2008)

Motif	GO ID	Enrichment Pvalue (-log10)	Description of the cluster
TGTCTW	GO:GNF0004	16.65	Sexual Development
TGTSTR	GO:0000278	16.65	Mitotic Cell Cycle
NTTGTCTN	GO:GNF0004	8.94	Sexual Development
GTGTRY	GO:0000278	11.30	Mitotic Cell Cycle
GTSTAN	GO:0000278	16.65	Mitotic Cell Cycle
TTKNTTG	GO:0005783	4.51	Endoplasmic Reticulum
ATSTST	GO:0000278	8.70	Mitotic Cell Cycle
TTTKTTK	GO:0005783	2.50	Endoplasmic Reticulum
TTNTTTKG	GO:0005783	3.11	Endoplasmic Reticulum
YCTAYA	GO:GNF0004	5.46	Sexual Development
NKTGTC	GO:GNF0004	16.65	Sexual Development
TTNTTGTY	GO:GNF0006	2.80	Sporozoite-Specific
AYACATM	GO:0005783	9.94	Endoplasmic Reticulum
TTNTTTTWG	GO:GNF0006	2.29	Sporozoite-Specific
TGWTCTA	GO:0006413	1.67	Translational Initiation
GTTSTM	GO:GNF0004	2.46	Sexual Development
TAKASA	GO:0005783	8.23	Endoplasmic Reticulum
GATAKATR	GO:0016052	3.63	Carbohydrate Catabolism
TWNTTTG	GO:0005783	6.41	Endoplasmic Reticulum
TRTACR	GO:0005783	16.65	Endoplasmic Reticulum
KACATWTAT	GO:0016052	5.54	Carbohydrate Catabolism
ACKAGC	GO:0006096	1.62	Glycolysis
YTMTGATT	GO:0006720	2.48	Isoprenoid Metabolism
TKTCTAWA	GO:0006413	2.01	Translational Initiation
ATAMTAMG	GO:0016052	2.10	Carbohydrate Catabolism
TTNGTY	GO:0005783	4.38	Endoplasmic Reticulum
TCTNTAWT	GO:0006457	2.52	Protein Folding
MATATACTA	GO:0000278	1.58	Mitotic Cell Cycle
WAWAATGA	GO:0005783	2.26	Endoplasmic Reticulum
AAACRR	GO:0006399	4.93	tRNA Metabolism
RTGTNG	GO:0006457	2.56	Protein Folding
TANAYAT	GO:0005783	9.46	Endoplasmic Reticulum
AMTTAGRT	GO:0005830	1.79	Cytosolic Ribosome (sensu Eukarya)
AAGMAATWA	GO:0019941	1.97	Modification-Dependent Protein Catabolism
TRYATATA	GO:0005783	8.86	Endoplasmic Reticulum
RCACATKTT	GO:0030894	1.86	Replisome
ARAGAGAAR	GO:0000278	3.91	Mitotic Cell Cycle
TYCTNTA	GO:0006457	4.68	Protein Folding

Table S2 Match of discovered motifs to those previously identified

Motif	Num in PF	Num in PY	Num in PF and PY	Z-score	Previously identified motif	Similarity score
TGTCTW	537	531	192	13.91	TGTCTW	1
TGTSTR	1492	1414	786	12.29	NTGTGTA	0.882
NTTGTCTN	325	323	93	10.80		
GTGTRY	1345	1305	648	9.78		
GTSTAN	1701	1654	945	8.91		
TTKNTTG	2536	2045	1585	6.35		
ATSTST	1578	1356	719	6.34		
TTTKTTK	2533	1984	1535	6.10		
TTNTTTKG	2198	1640	1135	5.93		
YCTAYA	1705	1505	834	5.84	TCTATAT	0.816
NKTGTC	1048	995	376	5.84		
TTNTTGTY	1601	1077	583	5.71		
AYACATM	2490	1765	1357	5.40		
TTNTTTTWG	1860	1182	714	5.26		
TGWTCTA	235	164	29	5.22	TGATCCA	0.751
GTTSTM	905	758	261	5.21		
TAKASA	2904	2480	2143	5.10	ATATAGA	0.816
GATAKATR	981	621	236	5.04		
TWNTTTG	2878	2444	2087	5.03		
TRTACR	2872	2403	2056	5.02	TATACAAC	0.764
KACATWTAT	1872	760	487	5.00		
ACKAGC	63	199	15	4.98		
YTMTGATT	267	228	39	4.93		
TKTCTAWA	557	459	116	4.93		
ATAMTAMG	362	311	60	4.85		
TTNGTY	2719	2382	1934	4.82		
TCTNTAWT	938	710	245	4.78		
MATATACTA	92	111	12	4.75		
WAWAATGA	1284	1005	440	4.74		
AAACRR	1846	2108	1205	4.73		
RTGTNG	1416	1275	596	4.68		
TANAYAT	2986	2357	2088	4.67		
AMTTAGRT	113	99	13	4.66		
AAGMAATWA	394	278	59	4.65		
TRYATATA	2671	1760	1428	4.62		
RCACATKTT	165	210	29	4.61		
ARAGAGAAR	241	200	36	4.60	GAGAGAA	0.865
TYCTNTA	1714	1230	678	4.60		

PF: P. Falciparum; FY: P. yoelii
Similarity score is between [0, 1].

Table S3 Enrichement of discovered motifs in functional gene sets (156 clusters reported by Zhou et al, 2008)

Motif	GO ID	Enrichment Pvalue (log10)	Description of the cluster
TGTCTW	GO:GNF0233	17.65	sexual development
TGTSTR	GO:GNF0211	17.65	gliding motility
NTTGTCTN	GO:PM16005087	9.45	The Plasmodium falciparum sexual development transcriptome: a microarray analysis using ontology-based pattern identification.|JOUR|Molecular & Biochemical Parasitology|2005
GTGTRY	GO:0005694	17.65	chromosome
GTSTAN	GO:GNF0211	17.65	gliding motility
TTKNTTG	GO:GNF0211	8.79	gliding motility
ATSTST	GO:GNF0233	11.23	sexual development
TTTKTTK	GO:GNF0004	8.58	Sexual Development
TTNTTTKG	GO:GNF0211	7.60	gliding motility
YCTAYA	GO:GNF0211	17.65	gliding motility
NKTGTC	GO:GNF0233	17.65	sexual development
TTNTTGTY	GO:0008092	6.64	cytoskeletal protein binding
AYACATM	GO:0015980	11.65	energy derivation by oxidation of organic compounds
TTNTTTTWG	GO:GNF0211	5.07	gliding motility
TGWTCTA	GO:0020008	3.80	rhoptry
GTTSTM	GO:GNF0211	9.28	gliding motility
TAKASA	GO:PM15032631	9.44	Mitochondrial and plastid functions as antimalarial drug targets.|JOUR|Current Drug Targets--Infectious Disorders|2004
GATAKATR	GO:0006413	6.09	Translational Initiation
TWNTTTG	GO:PM15032631	9.91	Mitochondrial and plastid functions as antimalarial drug targets.|JOUR|Current Drug Targets--Infectious Disorders|2004
TRTACR	GO:0016052	17.65	Carbohydrate Catabolism
KACATWTAT	GO:0005794	5.45	Golgi apparatus
ACKAGC	GO:PM14982620	3.28	Differential transcriptome profiling identifies Plasmodium genes encoding pre-erythrocytic stage-specific proteins.|JOUR|Molecular microbiology|2004
YTMTGATT	GO:PM12177071	4.00	Infectivity-associated changes in the transcriptional repertoire of the malaria parasite sporozoite stage.|JOUR|The Journal of biological chemistry|2002
TKTCTAWA	GO:GNF0211	5.62	gliding motility
ATAMTAMG	GO:0006090	3.87	pyruvate metabolism
TTNGTY	GO:GNF0211	17.65	gliding motility
TCTNTAWT	GO:GNF0004	4.06	Sexual Development
MATATACTA	GO:0004721	2.61	phosphoprotein phosphatase activity
WAWAATGA	GO:0006259	4.42	DNA metabolism
AAACRR	GO:PM15032631	7.60	Mitochondrial and plastid functions as antimalarial drug targets.|JOUR|Current Drug Targets--Infectious Disorders|2004
RTGTNG	GO:0008092	5.87	cytoskeletal protein binding
TANAYAT	GO:PM15032632	10.38	Computer assisted searches for drug targets with emphasis on malarial proteases and their inhibitors.|JOUR|Current Drug Targets--Infectious Disorders|2004
AMTTAGRT	GO:PM10395796	2.67	Interpolated Markov models for eukaryotic gene finding.|JOUR|Genomics|1999
AAGMAATWA	GO:0006090	3.90	pyruvate metabolism
TRYATATA	GO:GNF0207	9.21	drug resistance
RCACATKTT	GO:PM16940297	4.10	Interactions between merozoite surface proteins 1
ARAGAGAAR	GO:0044409	3.39	entry into host
TYCTNTA	GO:PM15935755	6.57	Proteome analysis of separated male and female gametocytes reveals novel sex-specific Plasmodium biology.|JOUR|Cell|2005

Table S4 Estimated β for k-mers of different size and different GC-content

GC number	k=6	k=7	k=8	k=9
0	0.06566	0.03447	0.01607	0.00696
1	0.01175	0.00467	0.00302	0.00544
2	0.00359	0.00605	0.01255	0.01416
3	0.01295	0.02865	0.03260	0.01754
4	0.01791	0.03920	0.06070	0
5	0	0	0	0
6	0	0	0	0
7		0	0	0
8			0	0
9				0

Table S5 Clusters of motifs discovered in the promoter regions of S. cerevisiae

	Expanded motif					K-mer					Match to known motif		
Motif	Num in species A	Num in species B	Num in A and B	Z-score	Motif	Num in species A	Num in species B	Num in A and B	Z-score	Transfac motif ID	Transfac motif consensus	Similarity score	Name
GCCCCCGG	22	24	14	40.86	GCCCCCGG	22	24	14	40.86	-	-	0	-
RAAAWTTTT	851	667	460	20.99	AAAAATTTT	368	297	176	12.73	-	-	0	-
CGGGTARN	398	377	277	20.20	CGGGTAAA	66	70	31	4.60	M00307	VTTACCCGB	0.941	REB1
RAAAAAAN	2709	2547	1946	20.11	AAAAAAAA	1431	1365	761	14.27	-	-	0	-
NTATATAN	2704	2561	1935	18.51	ATATATAT	741	750	282	8.18	M00713	YHYTTATAK	0.772	TBP
NCACGTG	471	529	288	17.58	TCACGTG	224	202	124	10.35	M00064	BVVCACGTKBKN	0.93	PHO4
GAYGANG	1668	1695	1026	17.34	GATGATG	474	460	160	6.09	-	-	0	-
ANGAYGA	2004	1988	1281	16.48	ATGATGA	628	491	194	5.75	-	-	0	-
GGNGGNG	1281	1534	782	16.36	GGTGGTG	257	316	100	6.96	-	-	0	-
GCNGCNG	1637	1886	1065	16.17	GCTGCTG	336	454	116	4.44	-	-	0	-
GGNGANG	1937	2083	1279	15.83	GGTGATG	280	276	84	5.04	-	-	0	-
GNGAYGA	1467	1430	818	15.13	GAGATGA	392	323	134	7.79	-	-	0	-
CGCGTNN	1039	1122	570	14.69	CGCGTCG	62	90	32	4.99	-	-	0	-
YTCTTYT	2959	2676	2098	14.56	CTCTTCT	610	642	199	4.19	-	-	0	-
CGSGNAA	1118	1090	601	14.48	CGCGAAA	174	184	81	7.45	M00307	VTTACCCGB	0.854	REB1
YGGTGGN	1185	1368	666	14.33	TGGTGGT	317	346	97	5.09	-	-	0	-
TNGGYGG	1044	1186	575	14.27	TCGGCGG	122	158	61	6.66	M00014	BDBTAGCCGCCVVVRVV	0.76	repressor
TGAAAAWN	2284	2071	1402	13.98	TGAAAATT	352	271	89	4.81	-	-	0	-
KTTTCNT	3507	3231	2800	13.88	TTTTCAT	1207	978	450	5.60	M00664	HTGAAAC	0.831	STE12
AARAAGR	3002	2803	2209	13.78	AAGAAGG	697	680	251	5.82	-	-	0	-
GNNTGGG	1504	1726	909	13.70	GAATGGG	226	202	75	6.57	M00213	DRMACCCABRCAYB	0.821	RAP1
GANGGNG	1824	1986	1147	13.68	GAAGGTG	326	359	102	4.09	-	-	0	-
ANYTCATC	985	926	486	13.61	ATCTCATC	182	183	85	7.81	-	-	0	-
NTCTTGY	2349	2268	1522	13.54	GTCTTGT	277	347	85	4.93	-	-	0	-
AARRGAA	3037	2703	2151	13.30	AAAGGAA	1095	913	381	4.43	-	-	0	-
ATGGNNG	2204	2188	1421	13.17	ATGGATG	299	293	85	4.28	-	-	0	-
NAGGGGN	1434	1547	820	13.01	AAGGGGA	261	283	81	4.49	M00308	VAGGGGV	0.991	STRE
NNAAATT	4084	3662	3516	12.68	TGAAATT	871	751	294	4.41	-	-	0	-
TTNTTYA	3731	3439	3096	12.48	TTGTTTA	657	618	206	4.66	-	-	0	-
TRGANGA	2110	1951	1248	12.16	TGGAAGA	588	541	188	5.18	-	-	0	-
GGYAARA	1481	1309	732	12.01	GGTAAGA	290	310	91	4.86	-	-	0	-
TNGGNTT	2679	2557	1834	11.89	TTGGATT	428	385	114	4.00	-	-	0	-
GANATNGA	1353	1276	673	11.85	GAAATGGA	140	141	44	4.41	-	-	0	-
ANRAAAT	3751	3373	3045	11.80	ATGAAAT	713	600	233	5.28	-	-	0	-
TGNNTGG	2265	2341	1502	11.77	TGTTTGG	344	368	104	4.93	M00213	DRMACCCABRCAYB	0.757	RAP1
TGGGNGN	1626	1853	985	11.76	TGGGAGA	202	213	60	4.54	-	-	0	-
GNYGGTG	1158	1337	617	11.75	GATGGTG	259	259	70	4.03	-	-	0	-
TTNAANGA	2062	1824	1148	11.73	TTCAAGA	261	250	70	4.81	-	-	0	-
TCRTTNA	2443	2133	1483	11.73	TCGTTTA	353	309	116	6.61	-	-	0	-
GANATGR	1923	1807	1098	11.73	GAAATGA	519	430	143	4.34	-	-	0	-
GTGGYRA	886	892	411	11.65	GTGGCAA	308	270	102	6.35	-	-	0	-
NATGGAN	2892	2655	2014	11.64	AATGGAT	419	342	122	5.53	-	-	0	-
CNCCACN	1769	2124	1160	11.62	CACCACG	139	219	49	4.02	-	-	0	-
TNGANCC	2041	2127	1297	11.57	TCGACCC	74	92	29	4.41	-	-	0	-
NYCTCGA	1205	1376	644	11.40	TTCTCGA	316	347	98	5.11	M00683	BHTCGAG	0.93	XBP1
NCCRTGG	999	1149	514	11.25	CCCATGG	130	155	55	5.78	-	-	0	-
TTRTTTNT	2105	1898	1180	11.20	TTGTTTAT	282	241	78	6.41	M01005	AGAACAAAGAAA	0.762	ste11
RCNAAGA	2289	2342	1514	11.17	ACCAAGA	453	590	169	5.08	M00275	WKRYBTCTTTGTTNNKYI	0.801	Mat1-Mc
AYGANTC	1324	1273	646	11.07	ATGACTC	201	189	66	5.63	M00204	VRRTGACTCH	0.825	GCN4
CCCYTNA	1196	1204	598	11.04	CCCTTAA	209	209	71	6.05	M00308	VAGGGGV	0.832	STRE
AACAANAR	1928	1858	1095	10.97	AACAATAG	197	232	61	4.87	M00728	YSYATTGTT	0.841	ROX1
TANAWAA	3656	3223	2859	10.95	TAAATAA	951	767	326	4.29	-	-	0	-
NYGGCCA	1160	1431	641	10.93	ATGGCCA	252	312	81	4.87	-	-	0	-
CTTGGYN	1641	1771	934	10.87	CTTGGCA	281	361	89	4.30	M01030	CCAAGVAM	0.878	Rim101p
GNTGYTG	1736	1873	1017	10.71	GCTGTTG	319	323	90	4.36	-	-	0	-
TGRNAAAA	1921	1754	1043	10.70	TGGCAAAA	213	171	60	4.27	-	-	0	-
TYGATRA	1692	1499	859	10.70	TTGATAA	697	499	192	4.32	-	-	0	-
TTTTCNAR	1693	1559	864	10.66	TTTTCAAA	533	477	147	5.07	-	-	0	-
NNAGATG	2871	2678	1999	10.65	AAAGATG	611	508	179	4.77	-	-	0	-
TTCATNR	2694	2355	1718	10.55	TTCATGA	271	248	84	5.83	M00664	HTGAAAC	0.773	STE12
ATGGNYA	1601	1549	835	10.35	ATGGTCA	233	241	68	4.23	-	-	0	-
TGRTCNA	1591	1628	862	10.12	TGGTCCA	185	226	64	5.50	-	-	0	-
TGGCNNC	2019	2163	1266	10.10	TGGCATC	242	257	73	4.03	-	-	0	-
ASCCGCN	841	909	407	10.04	AGCCGCC	196	220	99	7.72	M00014	BDBTAGCCGCCVVVRVV	0.838	repressor
AAAANGGN	2014	1823	1106	9.95	AAAATGGT	171	165	42	4.01	-	-	0	-
NGCTCRT	1200	1323	612	9.93	AGCTCAT	311	319	118	7.43	-	-	0	-
NGYGTCA	1100	1085	507	9.87	TGTGTCA	251	220	82	5.89	M00038	DBBBVTGACTCAYHBHB	0.761	GCN4
TTCCARR	1523	1564	803	9.78	TTCCAAA	702	665	230	4.95	-	-	0	-
TTCRATTS	549	479	200	9.75	TTCAATTC	199	179	57	5.24	-	-	0	-
TGGANRC	1518	1485	774	9.53	TGGAGAC	180	180	50	4.14	M00337	BRGCWTCCAYWWTK	0.826	GCR1
TCARNTC	1787	1728	966	9.46	TCAATTC	492	418	145	5.03	-	-	0	-
GCNANGG	1790	1924	1060	9.45	GCCAAGG	225	278	80	4.34	-	-	0	-
CNTTRGA	1758	1685	944	9.43	CTTTGGA	409	425	129	4.68	-	-	0	-
GGNCCAN	1628	1958	991	9.31	GGTCCAG	106	154	37	4.37	-	-	0	-
YTGNTGT	2145	2105	1281	9.27	TTGGTGT	348	339	93	4.01	M00030	TGATGTRVHB	0.804	MATa1
NTNTGTG	2611	2424	1673	9.26	GTTTGTG	270	290	76	4.22	-	-	0	-
CYATYGT	1179	1140	538	9.21	CTATTGT	499	434	154	5.70	M00276	YYHATTGTYD	0.843	Mat1-Mc
TNGATNT	3368	3158	2598	9.14	TGGATCT	229	252	63	4.09	-	-	0	-
TGGTNYC	1419	1494	733	9.09	TGGTTTC	375	416	111	4.04	-	-	0	-
NACATGR	1466	1441	741	9.09	TACATGA	286	251	87	5.19	M00031	BBVTGTHDWH	0.867	MATalpha2
ACNAMAA	2996	2852	2159	8.98	ACCACAA	315	350	99	4.37	-	-	0	-
TNTTNAAA	2868	2464	1821	8.93	TCTTCAAA	270	276	70	4.25	-	-	0	-
CCANTYG	1458	1476	735	8.60	CCAATTG	405	378	128	5.29	-	-	0	-

AWGTTYG	1098	1055	471	8.46	ATGTTTG	366	362	106	4.86	-	-	0	-
TGSTGCN	1469	1605	771	8.23	TGCTGCT	375	481	127	4.98	-	-	0	-
RANGGTAT	606	596	224	8.02	GAAGGTAT	102	96	36	4.72	-	-	0	-
SRGTTCGA	168	201	76	7.19	GGGTTCGA	42	55	27	5.97	-	-	0	-
GNTATNAA	1429	1178	575	6.96	GGTATGAA	67	53	20	4.13	-	-	0	-
GTYNGTAA	472	412	150	6.95	GTTAGTAA	72	62	19	4.84	-	-	0	-
TTTSGYGT	367	310	124	6.79	TTTCGTGT	94	93	33	4.58	-	-	0	-
TTGTGTMA	184	161	61	6.66	TTGTGTCA	95	85	38	5.47	-	-	0	-
YTTATCNG	538	500	176	5.92	TTTATCGG	51	57	20	4.27	-	-	0	-
RWTACGTA	277	266	78	4.85	ATTACGTA	84	90	26	4.16	-	-	0	-

Table S6 Clusters of motifs discovered in the promoter regions of C. elegans

	Expanded motif					K-mer					Match to known motif		
Motif	Num in species A	Num in species B	Num in A and B	Z-score	Motif	Num in species A	Num in species B	Num in A and B	Z-score	Transfac motif ID	Transfac motif consensus	Similarity score	Name
NYCTTATCA	1518	1706	578	20.44	TTCTTATCA	589	641	137	12.59	-	-	0	
CTGCGTCT	911	1111	374	19.44	CTGCGTCT	911	1111	374	19.44	-	-	0	
TGASTCAY	1139	1436	430	18.99	TGACTCAT	461	615	152	15.56	-	-	0	
NNATTTT	10890	10884	10881	18.39	CAATTTT	8228	7248	5623	5.81	-	-	0	
TGCGTCTC	857	1081	331	17.39	TGCGTCTC	857	1081	331	17.39	-	-	0	
YCCGCCY	2292	2803	1029	17.25	TCCGCCC	816	1393	294	13.04	-	-	0	
GTATAWAAG	362	344	104	16.97	GTATAAAAG	271	276	77	14.78	-	-	0	
NNATCGAT	5182	4731	2766	16.86	AAATCGAT	1695	1282	361	9.21	-	-	0	
TCTTCTYNT	3241	4747	1858	15.90	TCTTCTTCT	973	2087	370	11.40	-	-	0	
GAYGANGA	2889	3298	1291	15.78	GATGATGA	823	863	166	8.08	-	-	0	
TCTTYYTCT	2668	4080	1388	14.62	TCTTTTTCT	1261	1752	305	5.14	-	-	0	
TTTTNAN	10890	10879	10876	14.54	TTTTCAT	6952	6592	4353	4.08	-	-	0	
CYCCGCCY	1179	1668	482	14.52	CTCCGCCC	530	1049	193	9.89	-	-	0	
CYCTCTY	4441	5764	2819	14.32	CTCTCTC	2032	2843	803	9.14	-	-	0	
CGCGCNN	4334	4413	2216	14.23	CGCGCGG	351	416	63	5.76	-	-	0	
NCACGTG	3001	2046	970	14.14	CCACGTG	1257	932	284	9.09	-	-	0	
CTTYTYCTT	2439	3612	1168	14.08	CTTTTTCTT	1150	1553	259	5.17	-	-	0	
RTGACGTN	1991	1630	594	13.45	ATGACGTG	290	318	52	6.31	-	-	0	
GWNGATGA	2817	3082	1145	13.22	GTTGATGA	547	527	73	4.16	-	-	0	
TCTCTYNT	5227	6595	3563	13.08	TCTCTTCT	1267	2070	354	4.30	-	-	0	
YTCTTYTT	6926	7900	5340	12.77	TTCTTTTT	4523	5239	2453	9.14	-	-	0	
NTGACCTY	1581	1808	529	12.77	ATGACCTT	253	300	45	6.11	-	-	0	
TGCGTCW	1887	2028	658	12.74	TGCGTCA	633	583	103	4.06	-	-	0	
GATGAYRA	1955	2093	640	12.59	GATGACAA	492	461	85	7.57	M00229	BHWWGTCATBVH	0.798	
CTYYTTTT	7172	7852	5480	12.58	CTTTTTTT	5084	5558	2849	8.34	-	-	0	
ARGARGA	5381	6451	3553	12.55	AAGAAGA	3375	4213	1609	10.49	-	-	0	
GTGTTTGC	560	515	149	12.48	GTGTTTGC	560	515	149	12.48	-	-	0	
YCYTTCT	6037	7508	4532	12.38	TCCTTCT	2251	3337	906	6.39	-	-	0	
TACNGTAN	5101	6827	3545	12.16	TACAGTAA	1040	1470	212	4.22	-	-	0	
GAGYRTCGT	286	302	97	12.03	GAGTGTCGT	120	147	49	9.80	-	-	0	
GCNGCNG	5346	5057	2895	12.00	GCTGCTG	974	990	182	4.09	-	-	0	
YCTGCGT	1678	1755	540	11.91	TCTGCGT	1324	1490	409	11.54	-	-	0	
GCGYCTY	2791	3043	1129	11.71	GCGCCTT	894	1012	176	4.96	-	-	0	
CCNCCNC	5845	6835	4047	11.41	CCACCAC	1135	1309	268	5.35	-	-	0	
GANGATS	5722	6319	3684	11.39	GAAGATC	1386	1459	336	7.26	-	-	0	
GGCGCNN	4437	4713	2303	11.32	GGCGCCA	524	592	108	7.36	-	-	0	
TCGATYTT	3142	3118	1205	11.17	TCGATTTT	2856	2817	1019	10.89	-	-	0	
TCTTTTNT	6386	7158	4505	11.12	TCTTTTCT	1848	2511	580	5.98	-	-	0	
RTSAATCA	1942	1922	571	11.07	ATCAATCA	765	696	150	10.02	-	-	0	
CNCCCWC	4160	5148	2366	11.02	CACCCAC	679	713	139	5.66	-	-	0	
CTCNTCY	6859	8155	5439	10.98	CTCATCT	1558	1860	414	4.83	-	-	0	
YTTTTTTYT	7411	7392	5294	10.95	CTTTTTTTT	3105	3156	1048	5.31	-	-	0	
RTTTCGYAA	862	834	199	10.95	ATTTCGCAA	380	346	69	7.64	M00398	RTTACGTAAY	0.824	
GGRNGGA	3833	4818	2038	10.94	GGGAGGA	639	925	119	4.67	-	-	0	
ATGGAYGM	779	949	200	10.93	ATGGATGC	171	220	34	5.66	-	-	0	
TCNCTYTC	4013	4953	2208	10.92	TCACTTTC	1030	1078	210	5.81	-	-	0	
TTGCRTCAY	342	348	95	10.88	TTGCATCAT	162	182	44	9.89	-	-	0	
CAAGGWCR	595	652	153	10.84	CAAGGACA	213	236	39	5.95	-	-	0	
GANGWGG	4588	5280	2572	10.63	GAAGTGG	972	1139	203	5.69	-	-	0	
CGCNNCG	4545	4644	2320	10.62	CGCGACG	435	419	64	4.14	-	-	0	
GGAKCNG	3574	4036	1674	10.61	GGATCCG	479	587	80	4.31	-	-	0	
TTTYTCTY	7849	8366	6261	10.58	TTTTTCTT	5156	5483	2840	7.83	-	-	0	
GGRATGR	2646	3098	1044	10.55	GGAATGG	893	1098	218	8.65	-	-	0	
RACYGGTT	914	1032	235	10.52	AACCGGTT	259	297	65	9.36	M00951	WCYGGTTY	0.914	
GAYGGAN	5074	5392	2869	10.46	GATGGAT	1118	1267	256	6.99	-	-	0	
TTNTCYT	8018	8510	6488	10.43	TTTTCTT	4756	4811	2308	6.21	-	-	0	
CASCTGN	5119	5568	2999	10.38	CAGCTGA	1352	1577	346	6.28	-	-	0	
CTGGAAN	5785	6377	3716	10.29	CTGGAAA	3337	3710	1318	4.70	-	-	0	
NCCACGT	2961	2443	1000	10.29	GCCACGT	914	871	206	7.89	-	-	0	
TATGCWAAT	366	288	66	10.27	TATGCAAAT	269	194	43	8.56	-	-	0	
GRATGGS	2152	2608	776	10.25	GGATGGG	499	597	94	7.05	-	-	0	
NKCGTCG	3831	4385	1899	10.21	GTCGTCG	700	979	156	6.19	-	-	0	
CGAYRCTCC	325	374	120	10.15	CGACACTCC	142	180	51	5.49	-	-	0	
GCTGRRAA	2977	3341	1202	10.09	GCTGAAAA	1996	2312	636	9.52	-	-	0	
AMGATSA	4703	4814	2419	10.08	AAGATCA	1801	1609	403	5.49	-	-	0	
RGAARGA	5059	6050	3125	10.08	AGAAAGA	2538	3076	879	5.45	-	-	0	
CACCTGY	1440	1453	409	9.99	CACCTGC	697	657	143	7.24	-	-	0	
CGNYGGA	3772	4233	1809	9.98	CGCTGGA	571	642	96	4.14	-	-	0	
NGATCCR	4215	4768	2203	9.98	GGATCCA	728	946	148	5.00	-	-	0	
GAAYGGR	2941	3606	1274	9.96	GAATGGG	796	1007	182	8.13	-	-	0	
CGNGGAN	5563	5824	3315	9.87	CGTGGAG	690	633	116	4.78	-	-	0	
ATYGATNG	2709	2487	879	9.83	ATTGATCG	341	312	54	6.23	-	-	0	
GAAGARNC	3040	3588	1299	9.79	GAAGAATC	593	763	100	5.13	-	-	0	
GGATMNG	3349	3847	1501	9.77	GGATATG	763	890	149	5.78	-	-	0	
ANGGAGR	5028	5903	3047	9.75	AGGGAGG	439	547	67	4.40	-	-	0	
RNGATCG	3324	3736	1460	9.72	GAGATCG	622	690	105	4.27	-	-	0	
SGAGANC	4630	4850	2413	9.69	CGAGACC	1192	681	189	4.95	-	-	0	
TCTCNTT	7473	8206	5882	9.68	TCTCATT	3521	3750	1403	4.14	-	-	0	
ATCRANGA	3244	3224	1244	9.61	ATCAAGGA	381	385	66	7.14	-	-	0	
NGTYATCA	3067	3184	1186	9.53	CGTCATCA	465	545	80	4.30	M00229	BHWWGTCATBVH	0.782	
GRWTGGG	2029	2491	698	9.42	GGTTGGG	341	503	53	4.21	-	-	0	
YTNCTCTT	5050	6161	3174	9.35	TTGCTCTT	855	888	146	4.12	-	-	0	

TTGTTTRY	3283	3148	1220	9.30	TTGTTTAT	1335	1327	251	5.29	M00475	BBBTTGTTTACDHB	0.944
GRNATGG	4745	5176	2576	9.28	GATATGG	861	917	171	6.32	-	-	0
TKRCTCAT	1799	2023	550	9.27	TTGCTCAT	677	812	121	5.06	-	-	0
RNGAGCA	5574	5524	3156	9.20	GAGAGCA	1141	1250	246	5.55	-	-	0
GNCCNCC	4529	5722	2719	9.18	GACCACC	611	765	132	6.42	M01048	BGGGDGGT	0.758
YGAYATCA	1477	1734	413	9.11	TGATATCA	635	802	106	5.46	-	-	0
TCATNYC	7672	8016	5894	9.08	TCATTTC	4134	3936	1690	4.12	-	-	0
RGMAAGAA	2565	2796	883	9.01	GGAAAGAA	781	902	129	4.93	-	-	0
CCTYTTY	7274	8203	5724	9.00	CCTTTTT	4600	5352	2516	6.61	-	-	0
CANNTGAC	3603	4448	1785	8.99	CACATGAC	222	213	35	4.32	-	-	0
TGCWCTTY	2550	2582	874	8.97	TGCACTTT	888	852	168	6.25	-	-	0
CATCATNAT	862	973	192	8.91	CATCATCAT	439	455	77	6.31	M00230	VWDWATCATBHD	0.763
TNTRTTT	10726	10549	10407	8.90	TTTGTTT	6946	6129	4094	5.29	-	-	0
TGGRATGR	1150	1311	282	8.82	TGGGATGG	139	178	22	4.31	-	-	0
GCGSTCN	2611	2857	963	8.81	GCGGTCG	287	418	51	4.59	-	-	0
GCWCTYT	4721	5036	2525	8.80	GCACTTT	2013	1912	521	4.99	-	-	0
TGTKTAY	5377	5305	2931	8.78	TGTTTAT	2871	2901	935	5.21	M00475	BBBTTGTTTACDHB	0.865
KYTCTTT	9037	9315	7867	8.77	TTTCTTT	7062	7161	4829	5.96	-	-	0
TACAAACTA	258	204	40	8.76	TACAAACTA	258	204	40	8.76	-	-	0
GNCCWTC	3825	5078	2109	8.71	GTCCATC	723	874	144	4.71	-	-	0
NAARGGA	6235	7024	4281	8.71	AAAAGGA	2776	3273	1009	5.63	-	-	0
GGATNCN	6113	7129	4282	8.68	GGATTCT	1065	1690	270	5.05	-	-	0
YATTACAC	570	459	100	8.67	TATTACAC	362	275	50	6.41	-	-	0
GANYGATG	1806	1904	512	8.63	GATTGATG	415	395	68	7.01	-	-	0
YGTGGRT	2383	2365	765	8.62	CGTGGGT	266	245	42	5.21	-	-	0
GRTCAAGT	468	436	90	8.60	GATCAAGT	311	261	48	6.40	-	-	0
GGWSAAG	3115	3655	1324	8.60	GGTCAAG	513	587	97	6.29	-	-	0
RKGTGTG	3054	2896	1090	8.55	ATGTGTG	1475	1286	302	5.21	-	-	0
GCNCRTC	3793	3902	1687	8.54	GCGCATC	390	461	82	7.41	-	-	0
GACCWCN	3428	4206	1638	8.53	GACCACA	678	755	134	5.85	M01049	YGKGDGGT	0.759
ATGATNSA	2948	3087	1079	8.50	ATGATTGA	607	561	82	5.39	M00230	VWDWATCATBHD	0.849
CGTAAATM	893	602	157	8.50	CGTAAATC	618	428	111	8.24	-	-	0
YTGCGKAC	627	630	143	8.49	TTGCGGAC	200	230	44	5.57	-	-	0
TGGRWTC	3259	4360	1587	8.42	TGGAATC	1398	1904	356	4.06	-	-	0
GGYCTYG	2257	1968	655	8.41	GGTCTCG	1187	742	234	8.04	-	-	0
TYTCWCTT	4832	5443	2709	8.38	TTTCACTT	1777	1479	361	4.86	-	-	0
TCGWAAM	7141	7611	5219	8.33	TCGAAAA	5385	5905	3129	6.12	-	-	0
TNNTTGTT	9325	8712	7598	8.30	TTGTTGTT	1537	1405	304	4.86	-	-	0
RGAAGRT	4304	5051	2279	8.29	AGAAGGT	1083	1413	232	4.20	-	-	0
GRWCCAG	1877	2482	649	8.22	GAACCAG	663	844	121	4.25	-	-	0
NMTAATCC	1891	1829	511	8.18	ACTAATCC	186	193	28	4.56	-	-	0
TCYGGAN	6118	7645	4540	8.16	TCTGGAT	1239	1682	303	4.61	-	-	0
TNTNAAA	10884	10875	10866	8.14	TATAAAA	5243	4226	2237	5.27	-	-	0
ATCRTNATC	1127	1182	258	8.13	ATCATAATC	157	138	17	4.24	-	-	0
GGACCRS	1110	1613	318	8.09	GGACCAG	362	530	79	7.25	-	-	0
CMAGRAG	3721	4523	1835	8.08	CCAGGAG	549	739	108	5.66	-	-	0
TTTTTGRN	9674	9557	8580	8.04	TTTTTGAT	3583	4063	1484	4.76	-	-	0
AKAWGATA	1834	2069	506	8.03	ATAAGATA	431	529	57	5.48	M00230	VWDWATCATBHD	0.75
NAACTYAC	2959	2140	804	7.98	AAACTCAC	826	715	118	4.31	-	-	0
ATGAAKM	5945	5858	3469	7.95	ATGAATA	2619	2230	681	4.80	-	-	0
TYGRATTT	6046	5692	3405	7.95	TTGAATTT	3848	3397	1341	4.43	-	-	0
ACRAKGA	4250	4330	1973	7.93	ACAATGA	1928	1643	416	4.48	-	-	0
NTRTGAT	7685	7654	5617	7.91	TTGTGAT	2556	2389	718	4.34	-	-	0
GGRTTSG	1888	2343	613	7.89	GGGTTCG	358	452	58	4.37	-	-	0
GARGANA	8143	8727	6687	7.88	GAAGAAA	4366	4596	2023	4.93	-	-	0
CGRCCKC	1188	1449	317	7.86	CGACCGC	285	385	49	4.87	-	-	0
TTTNAAMA	10363	9692	9277	7.85	TTTCAAAA	6021	5508	3195	4.67	-	-	0
TAAYTTATT	1325	724	175	7.85	TAATTTATT	1086	589	120	6.44	-	-	0
GNAYGAT	5071	5176	2694	7.82	GCATGAT	743	689	115	4.28	M00230	VWDWATCATBHD	0.788
ATGGMWC	2951	3118	1109	7.82	ATGGCTC	835	906	163	5.18	-	-	0
ATCNWATC	2878	3040	1039	7.80	ATCAAATC	825	744	116	4.37	-	-	0
GNGGKCC	1592	2447	545	7.74	GTGGTCC	461	608	83	4.91	-	-	0
TGTAYATAN	1126	760	176	7.73	TGTACATAA	135	123	18	5.26	-	-	0
YTAAAGGC	719	594	123	7.71	TTAAAGGC	655	518	101	6.86	-	-	0
NAACANCA	5531	5491	3061	7.62	CAACAACA	763	834	133	4.94	-	-	0
ANTACAW	8456	7604	6093	7.62	ACTACAA	1821	1763	432	5.02	-	-	0
ATYRATGAG	514	472	88	7.60	ATTGATGAG	174	144	25	5.58	-	-	0
TTNCAGGA	2076	2663	712	7.56	TTTCAGGA	1118	1446	248	4.99	-	-	0
GGTCMCT	857	1179	217	7.53	GGTCACT	597	676	120	6.07	-	-	0
GARCAKGT	991	952	204	7.39	GAACAGGT	220	217	34	5.01	-	-	0
CCASGWC	1810	2106	559	7.39	CCAGGAC	368	449	64	4.83	-	-	0
ANACAACR	3174	3181	1152	7.38	AGACAACA	391	491	57	4.65	M00475	BBBTTGTTTACDHB	0.768
NAATWAA	10655	9965	9780	7.36	CAATAAA	4131	3118	1354	4.52	-	-	0
CCWTGRC	1850	2230	590	7.34	CCTTGAC	615	694	112	4.62	-	-	0
TGSCACG	1261	1197	293	7.34	TGGCACG	504	466	76	4.47	-	-	0
CMYATGT	3138	2885	1098	7.34	CACATGT	1107	1086	214	4.37	-	-	0
ACRYAAACA	596	536	104	7.33	ACATAACA	156	146	18	5.03	M00475	BBBTTGTTTACDHB	0.763
RAARAAC	9625	9400	8398	7.27	AAAGAAC	2069	2166	538	4.22	-	-	0
YTCCCAN	7015	7465	5057	7.21	TTCCCAT	2207	2384	664	4.50	-	-	0
YGTTRTTG	1716	1645	420	7.20	TGTTGTTG	795	756	128	5.09	-	-	0
TTTTWTG	9735	9439	8522	7.18	TTTTTTG	9147	8845	7545	7.12	-	-	0
ATCARTRT	2230	1744	517	7.13	ATCAATAT	1039	687	120	4.22	-	-	0
CATYTGNA	4437	4178	1961	7.12	CATCTGAA	909	1060	167	4.44	-	-	0
GACTAYNA	1658	1667	411	7.09	GACTACAA	439	435	64	5.06	-	-	0
TTYCATT	7947	7288	5519	7.08	TTTCATT	6433	5579	3482	4.87	-	-	0
TAWTTATTN	4460	2665	1284	7.05	TATTTATTG	513	301	48	5.37	-	-	0
ATAAAAGS	933	914	166	7.05	ATAAAAGG	444	512	64	5.75	-	-	0
TTTWTAW	10545	9940	9660	6.90	TTTTTAA	9024	7059	5953	4.32	-	-	0
ATCCKMTTA	229	239	41	6.85	ATCCGATTA	91	87	15	4.23	-	-	0

CCWGGAGW	593	707	130	6.84	CCAGGAGA	176	278	39	5.72	-	-		0
TTACRYAA	1515	1025	261	6.79	TTACACAA	547	406	66	4.88	M00398	RTTACGTAAY	0.936
TCRCACCT	384	441	82	6.61	TCACACCT	280	316	53	5.40	-	-		0
TYATYAAGG	247	217	42	6.61	TCATCAAGG	87	80	18	4.12	-	-		0
GRTGGNAT	1544	1656	384	6.58	GGTGGCAT	132	145	23	4.37	-	-		0
TATRMATA	2773	1824	620	6.57	TATAAATA	1390	832	180	5.10	-	-		0
TGGANRT	7500	7345	5257	6.50	TGGAAAT	4070	3703	1558	4.18	-	-		0
CTAARGAGG	43	63	16	6.50	CTAAAGAGG	36	48	14	6.45	-	-		0
RCATRATC	905	819	159	6.48	ACATAATC	355	297	38	4.08	-	-		0
ACCTTRR	2324	2312	695	6.37	ACCTTGG	439	499	81	5.89	-	-		0
YRATTAG	3671	3471	1398	6.32	CAATTAG	1276	1261	257	5.11	-	-		0
GARAWTT	10133	9820	9188	6.32	GAGATTT	2774	2867	895	4.82	-	-		0
CAWTYTT	10457	10156	9784	6.32	CATTTTT	8565	8020	6412	4.45	-	-		0
ARTCAYA	5034	4858	2486	6.26	AATCACA	2130	2081	541	4.10	-	-		0
TTYRGATGA	751	916	143	6.21	TTCAGATGA	300	422	43	4.31	-	-		0
GAATMNAT	4050	3826	1626	6.16	GAATACAT	533	464	61	4.07	-	-		0
CCRNAAA	9536	9640	8516	6.14	CCAAAAA	5963	5870	3390	4.58	-	-		0
GTGTCRKT	977	1171	217	6.11	GTGTCGTT	299	438	53	4.83	-	-		0
GTTGGGTTC	43	61	18	6.09	GTTGGGTTC	43	61	18	6.09	-	-		0
RTWTTTA	10157	9290	8726	6.00	ATTTTTA	8389	7079	5566	4.09	-	-		0
GTMATAAAA	552	518	70	5.93	GTCATAAAA	203	237	23	4.37	-	-		0
AGGGCCGYT	28	40	15	5.93	AGGGCCGTT	25	30	14	5.38	-	-		0
CGYAART	3366	2671	1056	5.90	CGTAAAT	1541	1090	279	5.24	-	-		0
CCCTWAAGA	54	60	17	5.78	CCCTAAAGA	40	44	15	5.74	-	-		0
ATAAAWAGA	640	581	77	5.74	ATAAATAGA	231	181	18	4.30	-	-		0
RCTWTACAC	194	184	34	5.73	ACTTTACAC	100	80	17	5.01	-	-		0
CAAANSA	8737	8664	7073	5.72	CAAATCA	2492	2359	696	4.29	-	-		0
CTTRGCG	768	735	141	5.67	CTTGGCG	578	568	91	4.13	-	-		0
NACASTT	8125	7741	5944	5.63	AACACTT	2398	2152	625	4.32	-	-		0
ATTACCTR	512	411	69	5.60	ATTACCTG	289	222	36	4.69	-	-		0
TTAYGATTC	226	273	33	5.59	TTATGATTC	160	196	19	4.14	-	-		0
TCATSRAAT	963	982	168	5.58	TCATCAAAT	404	317	45	4.75	-	-		0
MAKGACC	2033	2822	707	5.52	CAGGACC	322	424	53	4.13	-	-		0
CAAASTACA	434	434	64	5.50	CAAACTACA	321	295	44	5.26	-	-		0
STTGWTGAT	632	653	99	5.37	CTTGATGAT	123	132	16	4.15	-	-		0
TACGAGTM	222	200	34	5.33	TACGAGTC	81	87	14	4.13	-	-		0
GTAAWYATG	334	326	44	5.33	GTAAACATG	91	86	13	4.26	-	-		0
GGMRAGGT	553	553	92	5.31	GGAAAGGT	252	210	32	4.37	-	-		0
RANGAACAT	1130	1155	207	5.13	AAGGAACAT	137	182	19	4.57	-	-		0
GWACCWGT	630	745	116	5.09	GTACCAGT	122	154	22	4.43	M00951	WCYGGTTY	0.768
CACATGATT	114	118	19	5.08	CACATGATT	114	118	19	5.08	-	-		0
CAAGTTCGA	75	113	17	5.02	CAAGTTCGA	75	113	17	5.02	-	-		0
GGSCGTTGG	39	48	15	4.86	GGCCGTTGG	32	35	14	4.50	-	-		0
GGCCTARM	737	859	137	4.85	GGCCTAAA	363	391	55	4.17	-	-		0
CTTATGACA	56	72	11	4.79	CTTATGACA	56	72	11	4.79	-	-		0
CCAAAGGG	104	134	21	4.69	CCAAAGGG	104	134	21	4.69	-	-		0
TWCAKGTA	1511	1376	298	4.55	TTCATGTA	518	468	61	4.07	-	-		0
ACGATAAAC	76	72	12	4.54	ACGATAAAC	76	72	12	4.54	-	-		0
GGGTKCGGT	48	45	17	4.53	GGGTTCGGT	44	34	16	4.15	-	-		0
GAGCCAATG	63	75	18	4.52	GAGCCAATG	63	75	18	4.52	-	-		0
TTTGATCTA	505	390	55	4.48	TTTGATCTA	505	390	55	4.48	-	-		0
AGTGAAAGT	131	157	18	4.47	AGTGAAAGT	131	157	18	4.47	-	-		0
TATCAGWGA	257	333	35	4.34	TATCAGTGA	166	190	22	4.06	-	-		0
AGAAGCAAT	152	157	19	4.03	AGAAGCAAT	152	157	19	4.03	-	-		0

Table S7 Clusters of motifs discovered in the promoter regions of D. melanogaster

	Expanded motif					K-mer					Match to known motif			
Motif	Num in species A	Num in species B	Num in A and B	Z-score	Motif	Num in species A	Num in species B	Num in A and B	Z-score	Transfac motif ID	Transfac motif consensus	Similarity score	Name	
CAGCTGWT	1974	2102	754	18.19	CAGCTGTT	1198	1326	372	14.45	M00044	VBVABBTGHTDVBW	0.779	Sn
RCACRCAC	2600	3335	1192	16.51	GCACACAC	1049	1267	262	8.89	-	-	0	-
STATAWAAG	732	685	230	16.49	CTATAAAAG	254	272	73	10.11	M00252	BTATAWAWVBVVVBB	0.913	TATA
ATTTGCAT	1397	1371	425	15.99	ATTTGCAT	1397	1371	425	15.99	-	-	0	-
NATCGRTA	3532	3859	1634	15.01	TATCGATA	1742	1925	559	13.76	M00488	WVBTATCGATWDDH	0.892	DREF
TCCTCNTC	2247	2103	765	14.62	TCCTCCTC	998	961	235	9.85	-	-	0	-
NWGATAAG	2886	2569	1019	14.04	CTGATAAG	545	566	110	8.60	M00487	KNMTTATCVNH	0.915	mtTFA
CACGTGYN	1550	1534	473	13.68	CACGTGTT	213	210	37	6.03	JASPAR_MA0086	CAGGTG	0.762	sna
YAATTAGN	3322	2947	1258	13.63	CAATTAGC	560	492	115	9.87	M00696	YCAATTA	0.85	En
TGACGTCA	327	417	102	13.59	TGACGTCA	327	417	102	13.59	-	-	0	-
TTRCATAA	1524	1402	425	13.37	TTGCATAA	846	778	203	12.67	-	-	0	-
STCCTYG	4727	3902	2108	13.17	CTCCTCG	1992	1558	521	9.38	-	-	0	-
TTGTTGNY	5138	5457	2903	13.10	TTGTTGTT	1857	2154	540	8.47	-	-	0	-
CTYCTTGN	3505	2950	1313	12.76	CTCCTTGT	514	428	94	7.14	-	-	0	-
CAATTARN	4547	4213	2114	12.67	CAATTAAA	1582	1414	372	8.76	M00696	YCAATTA	0.926	En
YNACCTTG	2072	1699	598	12.66	TCACCTTG	444	337	63	4.77	-	-	0	-
CATAAATY	1746	1634	499	12.59	CATAAATT	1190	1081	269	9.86	-	-	0	-
TGYTGYTG	3605	4647	1862	12.59	TGTTGTTG	1417	1793	412	10.14	-	-	0	-
AGWGARAG	1713	3096	740	12.52	AGAGAGAG	683	1816	232	9.34	-	-	0	-
YTTAATTR	4413	3479	1763	12.25	TTTAATTG	1619	1397	378	8.95	M00020	VVDVYAATTAAG	0.938	Ftz
AGAGAGNG	1436	2855	606	12.23	AGAGACG	574	863	106	4.23	M00723	VSWGAGMGHVR	0.794	GAGA
NGTCCTK	5681	5079	2996	12.17	TGTCCTT	1620	1301	346	6.43	M00009	GGTCCTGC	0.818	Ttk
CCTYCWTG	1590	1209	406	12.13	CCTTCATG	311	244	54	6.63	-	-	0	-
TGATTGAY	928	1055	238	12.10	TGATTGAT	662	765	156	11.34	-	-	0	-
ACTYACC	2458	1792	692	11.96	ACTCACC	1403	1023	275	6.74	-	-	0	-
TCCTYGNT	2995	2536	1005	11.93	TCCTCGAT	524	415	71	4.73	-	-	0	-
RTGANTCA	2148	2193	689	11.92	ATGACTCA	232	242	35	5.22	M00199	VTGASTCAB	0.946	AP-1
GYCATNAA	3291	3111	1267	11.90	GCCATAAA	848	853	195	10.24	M00090	VBNTWDAYKVCNVB	0.939	Abd-B
ACYTGTTG	1057	1009	266	11.86	ACCTGTTG	413	411	94	9.63	M00044	VBVABBTGHTDVBW	0.779	Sn
CNTCATCN	4566	4460	2205	11.70	CATCATCA	794	895	168	7.88	-	-	0	-
CASNTCCT	3291	2474	1079	11.64	CAGCTCCT	935	728	173	6.73	-	-	0	-
CTCCTCNT	2035	1978	633	11.60	CTCCTCCT	1036	1024	227	7.91	-	-	0	-
NNAATAAN	9795	9255	8197	11.56	ACAATAAA	1481	1470	320	5.49	M00266	HWNVRTAAATABHNBV	0.788	Croc
RNCATGTG	2249	2122	711	11.50	GACATGTG	271	261	45	5.72	JASPAR_MA0086	CAGGTG	0.762	sna
CGATCNN	8196	7713	5910	11.43	CGATCTT	1194	924	199	4.20	-	-	0	-
CCTYCTTS	1868	1581	514	11.38	CCTTCTTC	633	560	96	4.55	-	-	0	-
STNCTCCA	3779	3253	1475	11.32	CTGCTCCA	947	717	164	6.39	-	-	0	-
CTYCTCCW	2767	2634	976	11.29	CTCCTCCA	995	864	196	6.97	-	-	0	-
YACCTGN	5939	5406	3262	11.25	TACCTGC	814	747	151	5.41	JASPAR_MA0086	CAGGTG	0.941	sna
TNCTCCTK	3432	3121	1304	11.24	TGCTCCTT	677	511	110	6.53	-	-	0	-
WTGCATAA	1402	1227	341	11.21	ATGCATAA	612	495	85	5.21	-	-	0	-
KGYTGCTG	2999	3862	1352	11.20	GGCTGCTG	698	959	131	4.66	-	-	0	-
NNTCCTK	8828	8351	6783	11.16	CATCCTG	1516	1008	262	4.24	M00009	GGTCCTGC	0.845	Ttk
GCGYGCN	4774	4950	2497	11.04	GCGCGCC	756	771	130	5.29	M00067	VVBBCACGCKVSHB	0.825	Hairy
RCCTTGNC	2236	2218	730	11.03	GCCTTGGC	504	527	85	5.18	-	-	0	-
CTNACCTK	1964	1690	539	11.02	CTCACCTG	446	315	70	6.35	-	-	0	-
AACCNGTT	1412	1358	372	10.96	AACCGGTT	333	353	72	8.79	M00951	WCYGGTTY	0.854	Grainyhead/Elf-1
TYRTCCTC	1529	1371	400	10.95	TTGTCCTC	430	356	62	4.69	-	-	0	-
CATGWCCW	1368	1156	337	10.93	CATGTCCA	510	502	106	8.31	-	-	0	-
CANCTSCT	3972	3508	1616	10.85	CATCTCCT	411	389	63	4.68	-	-	0	-
CTYSTCCT	2250	2090	699	10.82	CTTCTCCT	755	672	123	5.30	-	-	0	-
SCGCATGN	1667	1549	461	10.70	GCGCATGC	337	311	59	5.72	-	-	0	-
CGYATRCG	1099	1025	271	10.63	CGCATGCG	320	273	63	7.26	-	-	0	-
CRTAATTR	1861	1653	493	10.62	CATAATTA	776	693	135	6.97	M00020	VVDVYAATTAAG	0.84	Ftz
MAAACAAM	5376	5576	3007	10.62	AAAACAAA	3363	3287	1228	7.55	M00093	DVVDAAACWARDHBH	0.773	BR-C
SNTGCGC	5984	5366	3252	10.61	GCTGCGC	1717	1462	398	4.53	-	-	0	-
CCNSCCA	6781	7067	4599	10.58	CCACCCA	1723	1619	401	4.53	-	-	0	-
TGACMTTG	762	675	166	10.56	TGACATTG	441	421	84	8.45	-	-	0	-
TCNTYGTC	2970	3211	1170	10.55	TCATTGTC	378	453	62	5.39	-	-	0	-
TGGCCNAR	3998	3370	1564	10.54	TGGCCGAG	527	506	100	7.14	-	-	0	-
TSNTCCAG	3222	2839	1147	10.46	TCCTCCAG	816	613	131	5.21	-	-	0	-
GTGGCCAN	2692	1983	779	10.46	GTGGCCAT	633	511	102	5.05	-	-	0	-
ATYGATTTY	1430	1232	341	10.45	ATTGATTT	579	547	110	9.35	-	-	0	-
RCCTTSTC	1150	1147	291	10.42	GCCTTCTC	375	385	63	5.11	-	-	0	-
CAAGGNCR	2243	2119	697	10.39	CAAGGCCA	496	444	76	5.13	-	-	0	-
TGMTGNAG	3353	3203	1282	10.39	TGCTGGAG	628	611	108	6.15	-	-	0	-
CCWGCNG	7792	7532	5508	10.38	CCAGCAG	2583	2843	854	4.93	-	-	0	-
GCNGCNGC	6477	7479	4596	10.35	GCAGCTGC	1260	1347	294	6.78	M00171	SVGCYGYBGYNGCVKV	0.785	Adf-1
TSWCCTC	3873	3716	1648	10.32	TCACCTC	1043	915	202	5.81	-	-	0	-
GNTGCWGC	4477	5049	2358	10.28	GCTGCAGC	1041	1239	215	4.59	-	-	0	-
TGCRTCAY	1047	1048	255	10.25	TGCATCAC	283	252	40	4.36	-	-	0	-
YGCACNTG	2915	2513	965	10.25	TGCACTTG	779	689	141	6.72	M00060	VHRRCAGGTGYVV	0.819	Sn
CAYATGY	5182	5532	2900	10.19	CATATGT	2304	2895	774	5.87	-	-	0	-
AARTGCAN	5410	5120	2813	10.15	AAATGCAT	1169	1130	241	7.59	-	-	0	-
GCWNCTGC	4525	4911	2329	10.13	GCTCCTGC	685	649	105	4.05	M00009	GGTCCTGC	0.751	Ttk
CTGYTSC	6290	6747	4103	10.13	CTGCTGC	3271	3974	1423	7.58	-	-	0	-
CTYCTTNA	3210	2630	1051	10.12	CTTCTTCA	807	625	130	6.46	-	-	0	-
ATGTANRT	4865	5092	2517	10.10	ATGTACAT	1246	1839	315	5.89	-	-	0	-
TCNCKCTC	2436	3711	1078	10.07	TCGCGCTC	303	301	46	4.41	M00723	VSWGAGMGHVR	0.879	GAGA
TGCAMYTT	2854	2463	904	10.06	TGCACTTT	1018	844	163	5.02	-	-	0	-
ATGAYSTC	1032	960	234	10.05	ATGATCTC	330	308	45	4.67	-	-	0	-
CWGCTCC	4150	3854	1789	9.96	CTGCTCC	2592	2259	761	7.33	-	-	0	-

NTGACNCA	3502	3213	1331	9.94	ATGACACA	261	290	38	4.86	M00199	VTGASTCAB	0.86	AP-1
CATGWCMA	1392	1347	354	9.92	CATGTCAA	268	338	40	4.62	-	-	0	-
GATSANC	6376	5741	3612	9.90	GATGAGC	1456	1226	296	5.19	-	-	0	-
CRTAMTTG	1355	1169	311	9.88	CGTACTTG	294	234	50	6.58	M00696	YCAATTA	0.786	En
GMCATNTT	3462	3282	1316	9.88	GCCATTTT	949	873	173	6.30	-	-	0	-
AANTRAAA	8519	7853	6177	9.86	AATTGAAA	1924	1708	466	7.28	-	-	0	-
TGTANRTA	4814	4991	2447	9.85	TGTACATA	1284	1803	325	6.37	-	-	0	-
GCNGCCAT	1781	1952	543	9.82	GCTGCCAT	609	666	107	5.30	-	-	0	-
TATAAATAG	322	324	69	9.80	TATAAATAG	322	324	69	9.80	M00252	BTATAWAWVBVVVBB	0.829	TATA
GMTCSTC	4131	3656	1709	9.80	GATCCTC	1197	851	194	4.01	-	-	0	-
ACYTTSAC	1361	1243	324	9.79	ACTTTCAC	487	409	84	8.12	-	-	0	-
NNCGCGC	6472	6479	4059	9.78	CTCGCGC	795	801	142	4.40	-	-	0	-
SKCCAGC	5599	5368	3046	9.78	GTCCAGC	1518	1440	336	4.49	-	-	0	-
KTCTTCWT	2328	2272	706	9.77	TTCTTCAT	682	603	94	4.64	-	-	0	-
NAWGGCCA	3785	3388	1478	9.77	GAAGGCCA	446	429	69	4.50	-	-	0	-
AAAMNAAA	9104	8847	7319	9.74	AAACCAAA	1588	1646	344	4.40	M00093	DVVDAAACWARDHBH	0.789	BR-C
YKCACCT	3972	3567	1618	9.74	CGCACCT	890	804	170	5.69	JASPAR_MA0086	CAGGTG	0.853	sna
CGNCGSC	6216	6061	3690	9.72	CGCCGCC	1820	1744	419	4.13	-	-	0	-
NYGTTGAC	2129	2124	650	9.70	TCGTTGAC	356	317	60	6.15	-	-	0	-
CNTCRTAG	1426	1336	354	9.70	CATCGTAG	274	242	40	4.46	-	-	0	-
CASNTGC	8180	7671	5840	9.69	CACTTGC	1788	1532	391	4.62	M00060	VHRRCAGGTGYVV	0.779	Sn
TCNTYCAT	2795	3130	1057	9.68	TCGTTCAT	322	385	53	5.96	-	-	0	-
CCTGNAG	3748	3731	1589	9.65	CCTGCAG	1804	1678	454	6.38	-	-	0	-
TNCTYGTT	3412	3228	1271	9.65	TGCTTGTT	588	517	83	5.12	-	-	0	-
GCTGNYG	8094	8395	6260	9.62	GCTGATG	1784	1710	430	5.62	-	-	0	-
TGAYCTY	3934	3540	1578	9.62	TGATCT	1585	1434	343	5.70	M00111	GRGGTCABB	0.756	CF1
CATGNTNG	3232	3008	1158	9.57	CATGATGG	261	268	43	5.81	-	-	0	-
NTGATTKAT	1625	1680	446	9.56	TTGATTTAT	459	458	74	6.26	-	-	0	-
TGATGYS	3937	3982	1735	9.50	TGATGCG	1028	1064	224	7.06	-	-	0	-
ATGTCRYC	1082	1001	245	9.50	ATGTCATC	349	343	50	4.80	-	-	0	-
CGCWKCAT	1224	1287	313	9.50	CGCAGCAT	371	363	66	6.24	-	-	0	-
RTTGATK	6095	6301	3734	9.47	ATTGATG	1739	2057	456	4.93	-	-	0	-
CCRGNTC	6174	5606	3435	9.45	CCAGCTC	1973	1896	530	4.45	-	-	0	-
YSTCCAG	4579	4420	2170	9.44	TGTCCAG	1628	1484	400	7.41	-	-	0	-
GCTCCWN	7296	6887	4772	9.42	GCTCCTG	1639	1440	391	7.28	-	-	0	-
NCGCWGCA	3209	3207	1231	9.42	TCGCTGCA	474	497	81	5.43	-	-	0	-
TYGWTGC	5537	5710	3143	9.39	TTGATGC	1425	1468	312	5.08	-	-	0	-
CCACRTNG	2388	2076	704	9.38	CCACATAG	319	335	48	5.19	-	-	0	-
YGATGACR	1381	1336	337	9.36	TGATGACA	324	323	41	4.00	-	-	0	-
KCRCGCA	3364	3379	1346	9.36	TCGCGCA	944	844	183	5.77	-	-	0	-
GSNCCTG	5524	5859	3224	9.35	GCGCCTG	1161	1093	226	4.49	M00009	GGTCCTGC	0.873	Ttk
TTWAWTGC	3177	2742	1048	9.34	TTTATTGC	963	922	159	5.15	M00090	VBNTWDAYKVCNVB	0.785	Abd-B
GCWCCWG	4786	4622	2333	9.33	GCACCAG	1374	1594	332	4.78	JASPAR_MA0086	CAGGTG	0.793	sna
CNKCGTC	5646	5603	3162	9.31	CGTCGTC	1216	1308	264	4.92	-	-	0	-
TTNCAYTT	6777	6171	4008	9.30	TTTCATT	2145	1923	515	5.22	-	-	0	-
TSCAGNTC	3612	3294	1382	9.30	TCCAGCTC	771	774	154	6.98	-	-	0	-
CCTTGWAS	1230	1137	290	9.29	CCTTGAAC	371	359	54	4.22	-	-	0	-
GCRTGCGY	1028	1037	245	9.29	GCATGCGC	328	295	56	5.79	-	-	0	-
TTSACCTY	1412	1346	347	9.29	TTGACCTC	260	215	36	4.45	-	-	0	-
ANAAWCAA	6813	7030	4534	9.28	ACAATCAA	609	821	100	4.33	JASPAR_MA0010	VBVAWWRACAARWN	0.782	Broad-complex_
GNGTGKG	6298	7263	4347	9.26	GCGTGGG	896	876	169	5.15	-	-	0	-
TNATGAYG	2394	2337	739	9.26	TGATGATG	773	837	150	7.09	-	-	0	-
GYATRCGC	1024	968	235	9.25	GCATACGC	284	295	50	5.65	-	-	0	-
TTGTNGTY	4284	4475	2008	9.25	TTGTAGTC	323	305	43	4.44	-	-	0	-
ATTKNAAT	7684	6773	4894	9.21	ATTGAAAT	1575	1432	310	4.16	-	-	0	-
GTGAARKT	1717	1607	443	9.19	GTGAAATT	744	720	145	8.39	-	-	0	-
AAAGCTYN	3336	2866	1135	9.15	AAAGCTTT	1041	792	164	5.74	-	-	0	-
AACNGCAR	3960	4412	1874	9.14	AACCGCAA	385	431	70	6.05	-	-	0	-
GYGTSAC	2909	2819	1021	9.12	GTGTGAC	1312	1269	271	5.64	-	-	0	-
GCSCACNT	2557	2205	756	9.10	GCCCACAT	437	639	66	4.04	-	-	0	-
AWWTCAATT	2074	2052	588	9.06	AATTCAATT	686	655	117	7.24	-	-	0	-
CTTGWYG	4411	4132	1961	9.05	CTTGATG	1548	1248	326	6.51	-	-	0	-
SKCCATG	4101	3960	1791	9.05	GTCCATG	1150	1062	222	4.76	-	-	0	-
ACWTGACN	1913	1893	533	9.03	ACTTGACC	309	272	45	4.69	-	-	0	-
GACANYTG	2592	2631	872	9.03	GACATTTG	478	460	67	4.62	-	-	0	-
CCTYGCGN	1544	1364	382	9.02	CCTTGCGC	344	247	50	4.90	-	-	0	-
TGNYGCC	7578	7497	5315	8.96	TGATGCC	1219	1231	264	5.97	-	-	0	-
TGWCANTT	3173	3147	1166	8.94	TGACAATT	555	516	80	5.09	-	-	0	-
GCTRATNA	3194	2922	1111	8.94	GCTGATGA	542	501	81	4.85	-	-	0	-
NTGNTGAT	5511	5295	2901	8.92	ATGCTGAT	534	592	81	4.88	-	-	0	-
GGCNTNCT	3744	3345	1425	8.88	GGCCTTCT	416	355	70	5.89	-	-	0	-
TCCACNTS	3174	2840	1096	8.84	TCCACATG	372	376	56	4.55	-	-	0	-
YGGCGSC	4897	4483	2288	8.83	CGGCGGC	1834	1846	480	7.01	-	-	0	-
TCRWAGTC	1327	1349	322	8.81	TCAAAGTC	534	590	79	4.43	-	-	0	-
NTTNATGT	6462	5837	3642	8.80	CTTGATGT	447	336	58	4.57	M00710	WCATTWAM	0.763	Zen
CCARGGY	3378	3710	1433	8.80	CCAAGGC	1181	1404	262	4.08	-	-	0	-
GNCMATAA	3003	2673	984	8.79	GGCCATAA	576	500	90	5.30	-	-	0	-
GACATNRT	2098	1930	574	8.79	GACATAAT	296	256	36	4.35	-	-	0	-
AKGCCNTC	1968	1974	564	8.78	ATGCCCTC	295	270	44	4.57	-	-	0	-
NSGGCGT	5201	4530	2451	8.78	AGGGCGT	784	664	136	4.90	-	-	0	-
NRTCACGT	1296	1199	299	8.77	AGTCACGT	164	177	26	4.68	-	-	0	-
GCCNGCN	9421	9166	7807	8.77	GCCAGCA	2448	2574	783	6.25	-	-	0	-
GCRTAMTT	1583	1297	358	8.72	GCATACTT	448	333	64	5.53	-	-	0	-
NTTGCGCN	4129	3488	1603	8.69	CTTGCGCT	463	336	67	4.70	-	-	0	-
RAAMGTGA	2018	2068	585	8.69	GAAAGTGA	629	653	99	4.75	-	-	0	-
ANGYGTG	6147	6413	3813	8.68	AGGTGTG	922	995	171	4.10	-	-	0	-
GCAAWTTR	2987	2615	955	8.67	GCAAATTG	874	850	153	4.62	-	-	0	-
SGCCAGYT	2009	1469	481	8.59	GGCCAGCT	717	534	107	4.62	-	-	0	-
CRAAMAA	8124	8346	6226	8.59	CAAAAAA	4885	5378	2601	7.14	-	-	0	-

TTAYGWCA	975	887	196	8.57	TTATGACA	369	327	50	5.28	-	-	0	-
AWGTTGCM	1862	1681	478	8.54	ATGTTGCC	461	478	65	4.11	-	-	0	-
CTTGTMST	1466	1171	321	8.50	CTTGTACT	382	339	52	4.57	-	-	0	-
TGNCCTY	6708	6282	4061	8.50	TGCCCTC	1106	1121	218	4.10	-	-	0	-
YGCGCKC	2913	2994	1060	8.45	TGCGCTC	1248	1120	249	4.94	-	-	0	-
KCNGCCT	6260	6693	4020	8.44	GCCGCCT	1490	1534	333	4.12	-	-	0	-
TCWYCTC	4279	4627	2081	8.41	TCATCTC	1127	1268	229	4.08	-	-	0	-
GCCACNNC	4910	5184	2576	8.40	GCCACGCC	623	453	92	4.60	-	-	0	-
ACAAARS	6418	6717	4117	8.39	ACAAAAC	2898	3104	1008	5.80	-	-	0	-
KCATGNA	6254	6600	3959	8.39	TCATGAA	1222	1163	236	4.79	-	-	0	-
TTGNSCAG	3928	3271	1460	8.37	TTGCCCAG	598	427	80	4.28	-	-	0	-
CGCGNCW	4510	4694	2219	8.34	CGCGCCA	917	956	176	4.69	M00067	VVBBCACGCKVSHB	0.76	Hairy
YNGGCGC	5766	5235	3019	8.28	TTGGCGC	1671	1486	375	4.82	-	-	0	-
TANTKTT	10370	9870	9146	8.26	TATTTTT	6547	6002	3694	4.05	-	-	0	-
GTRATTK	5617	5152	2891	8.26	GTGATTT	1888	1568	402	4.31	-	-	0	-
ATGTCNWG	2096	2013	589	8.23	ATGTCGTG	213	223	31	4.34	-	-	0	-
NGTCMTG	5193	4903	2592	8.22	TGTCATG	775	892	135	4.12	M00009	GGTCCTGC	0.818	Ttk
WGYACTT	6143	5371	3241	8.20	TGTACTT	1978	1713	459	5.21	-	-	0	-
ATWAATTRC	1033	1000	221	8.16	ATTAATTGC	277	237	40	4.36	M00455	DHRATTAVDN	0.75	dri
TGTTTGNY	4611	4786	2251	8.16	TGTTTGCT	882	907	143	4.16	-	-	0	-
GTACNTSA	1538	1569	390	8.14	GTACATCA	238	270	34	4.31	-	-	0	-
GGATSCN	6123	5124	3117	8.14	GGATGCT	1063	1017	195	4.30	-	-	0	-
ATGCANG	3808	3891	1619	8.13	ATGCAGG	1199	1083	240	5.66	-	-	0	-
TCGATNA	5288	4993	2656	8.13	TCGATCA	1219	1193	238	4.54	-	-	0	-
WTGACAMA	1781	1892	479	8.12	TTGACACA	357	345	50	4.64	-	-	0	-
ATTARCRC	1148	989	235	8.08	ATTAACAC	389	354	58	5.93	JASPAR_MA0094	TAAT	0.771	Ubx
CAAWTAANT	2666	2265	766	8.07	CAAATAAAT	732	620	97	4.12	M00020	VVDVYAATTAAG	0.785	Ftz
CATGYCR	3402	3756	1433	8.07	CATGCCA	1254	1597	310	5.66	-	-	0	-
AANGCCAW	4086	4044	1767	8.05	AAGGCCAA	607	572	96	4.93	-	-	0	-
GGCCNYC	6040	5819	3435	8.04	GGCCTTC	1104	1032	216	4.86	-	-	0	-
CGANGAG	4327	4064	1886	8.02	CGAGGAG	1761	1479	408	6.13	-	-	0	-
TRYACAT	6407	6597	4021	7.99	TGTACAT	2435	2924	836	6.79	-	-	0	-
TTATGRS	4034	3778	1654	7.97	TTATGAC	1089	868	184	4.96	M00090	VBNTWDAYKVCNVB	0.81	Abd-B
RCWAATGA	2043	1841	529	7.97	ACAAATGA	734	726	112	5.00	-	-	0	-
NGCCAWTT	4225	3812	1731	7.94	CGCCATTT	536	574	88	4.71	M00696	YCAATTA	0.752	En
CTTNAKCT	3091	2717	999	7.90	CTTCAGCT	710	604	102	4.09	-	-	0	-
CCWGCWTG	1506	1356	362	7.87	CCAGCATG	401	389	68	5.16	-	-	0	-
TNATTKT	10471	10138	9465	7.85	TTATTTT	6292	5463	3272	4.32	-	-	0	-
CATTRTCN	2638	2743	880	7.83	CATTGTCT	365	411	56	5.00	-	-	0	-
GTNGNCGT	3701	3561	1456	7.82	GTGGGCGT	614	433	101	5.96	-	-	0	-
AWGTMCTT	2005	1703	490	7.79	AAGTACTT	557	522	80	5.12	-	-	0	-
AWGTTSTC	1555	1476	366	7.71	AAGTTCTC	457	386	65	5.51	-	-	0	-
KGNTCGTA	1366	1418	326	7.66	GGATCGTA	226	208	33	4.40	-	-	0	-
ASGTGNT	5667	4847	2762	7.65	AGGTGTT	1152	944	197	4.38	M00044	VBVABBTGHTDVBW	0.765	Sn
CGGANGY	5213	4539	2430	7.62	CGGAAGT	1039	741	169	4.67	JASPAR_MA0026	BMGGAAR	0.781	Eip74EF
TSCTNGC	7088	6625	4443	7.60	TGCTTGC	1490	1509	323	4.14	-	-	0	-
KTCWACAA	1810	2044	499	7.56	TTCAACAA	674	844	111	4.93	-	-	0	-
SCCNAAA	7986	7933	5830	7.49	CCCAAAA	2765	2736	861	4.58	-	-	0	-
ANAAWAG	9286	9376	7844	7.48	ATAAAAG	3034	3153	1060	4.59	-	-	0	-
RCNTGCCA	2531	2758	862	7.45	GCATGCCA	377	467	60	4.49	-	-	0	-
CASYTCAT	1557	1299	339	7.45	CAGTTCAT	485	402	65	4.52	-	-	0	-
GMNCTCCT	2387	2117	679	7.43	GAACTCCT	423	314	58	4.23	-	-	0	-
AAWTANTT	8462	7419	5770	7.39	AAATATTT	3797	2927	1189	4.40	-	-	0	-
TGACAGSN	1835	1927	502	7.35	TGACAGCT	410	400	60	4.54	-	-	0	-
ARCGSAAG	1296	1292	297	7.25	AACGGAAG	360	396	57	4.93	-	-	0	-
TTGCNCT	5267	4953	2610	7.25	TTGCCCT	1334	1234	270	4.37	-	-	0	-
AAYTTKC	7015	6710	4422	7.19	AACTTTC	2083	1688	468	4.97	-	-	0	-
TTTCACAM	1340	1175	268	7.17	TTTCACAC	568	543	81	4.60	-	-	0	-
TCATTARR	1815	1521	409	7.15	TCATTAAA	990	776	166	6.62	M00018	BDVBVBBTAATKRBBHHVI	0.786	Ubx
KGTCAATN	2338	2290	680	7.15	GGTCAATG	312	295	44	4.08	-	-	0	-
GACACSNT	1364	1461	331	7.15	GACACCAT	264	275	38	4.02	-	-	0	-
GTWGTNG	6486	6574	4043	7.08	GTAGTAG	846	825	143	4.16	-	-	0	-
TMGTASAT	1068	1191	227	7.07	TCGTAGAT	304	312	43	4.56	-	-	0	-
CNATTGTN	5117	5009	2547	7.03	CCATTGTT	599	585	89	4.24	-	-	0	-
CANSTTTT	4291	3911	1746	7.03	CATCTTTT	552	555	75	4.39	-	-	0	-
CCATYCA	2841	3232	1062	7.00	CCATCCA	1514	1761	375	4.64	-	-	0	-
WGGAGTAY	1072	1025	215	7.00	AGGAGTAC	290	299	46	5.09	-	-	0	-
GCAYGGN	5236	5609	2900	6.96	GCATGGT	806	728	130	4.07	-	-	0	-
CCRTSAT	3197	3281	1193	6.89	CCATGAT	980	984	178	4.46	-	-	0	-
ATAAWTTAY	1868	1437	395	6.87	ATAAATTAT	726	552	97	4.65	-	-	0	-
ATGTSAYA	1206	1126	242	6.86	ATGTGATA	372	303	47	4.98	-	-	0	-
GYCTGGN	5900	6282	3556	6.85	GTCTGGT	801	1031	153	4.18	-	-	0	-
GTGTGKMA	1481	1855	399	6.66	GTGTGGAA	426	492	64	4.22	-	-	0	-
TYCTNGAA	3138	2656	964	6.54	TTCTGGAA	511	513	72	4.39	-	-	0	-
YGTNGGC	6627	6484	4072	6.46	TGTAGGC	815	782	138	4.07	-	-	0	-
CTKTSATT	2008	2080	534	6.38	CTGTGATT	387	397	57	5.36	-	-	0	-
ASTACATC	450	419	71	6.35	AGTACATC	260	232	36	4.96	-	-	0	-
CRWTGAAC	1358	1361	301	6.35	CAATGAAC	429	392	62	5.23	-	-	0	-
ATSGAGKA	1349	1365	298	6.22	ATGGAGTA	263	248	35	4.45	-	-	0	-
CSCACCM	3582	3951	1526	6.20	CCCACCA	1500	1524	359	5.47	-	-	0	-
CARCATNA	3110	3317	1139	6.07	CAACATAA	557	490	72	4.21	-	-	0	-
CGTAGGTG	225	158	36	6.00	CGTAGGTG	225	158	36	6.00	-	-	0	-
CMGYCCAT	1081	1066	219	5.99	CCGCCCAT	328	270	54	5.51	-	-	0	-
TAAAMGYC	1348	1203	269	5.89	TAAAAGCC	537	501	78	4.40	-	-	0	-
GACCCANWT	631	563	129	5.83	GACCCAATT	88	84	20	4.27	-	-	0	-
AGAWGATCA	240	224	46	5.80	AGATGATCA	104	103	20	4.28	-	-	0	-
WWTGATTAA	1093	964	188	5.61	AATGATTAA	277	234	35	4.71	M00455	DHRATTAVDN	0.817	dri
AAGSCMTA	996	1058	191	5.30	AAGCCATA	404	402	59	4.66	-	-	0	-
CGTAGTA	588	544	96	5.12	CGTAGTA	588	544	96	5.12	-	-	0	-