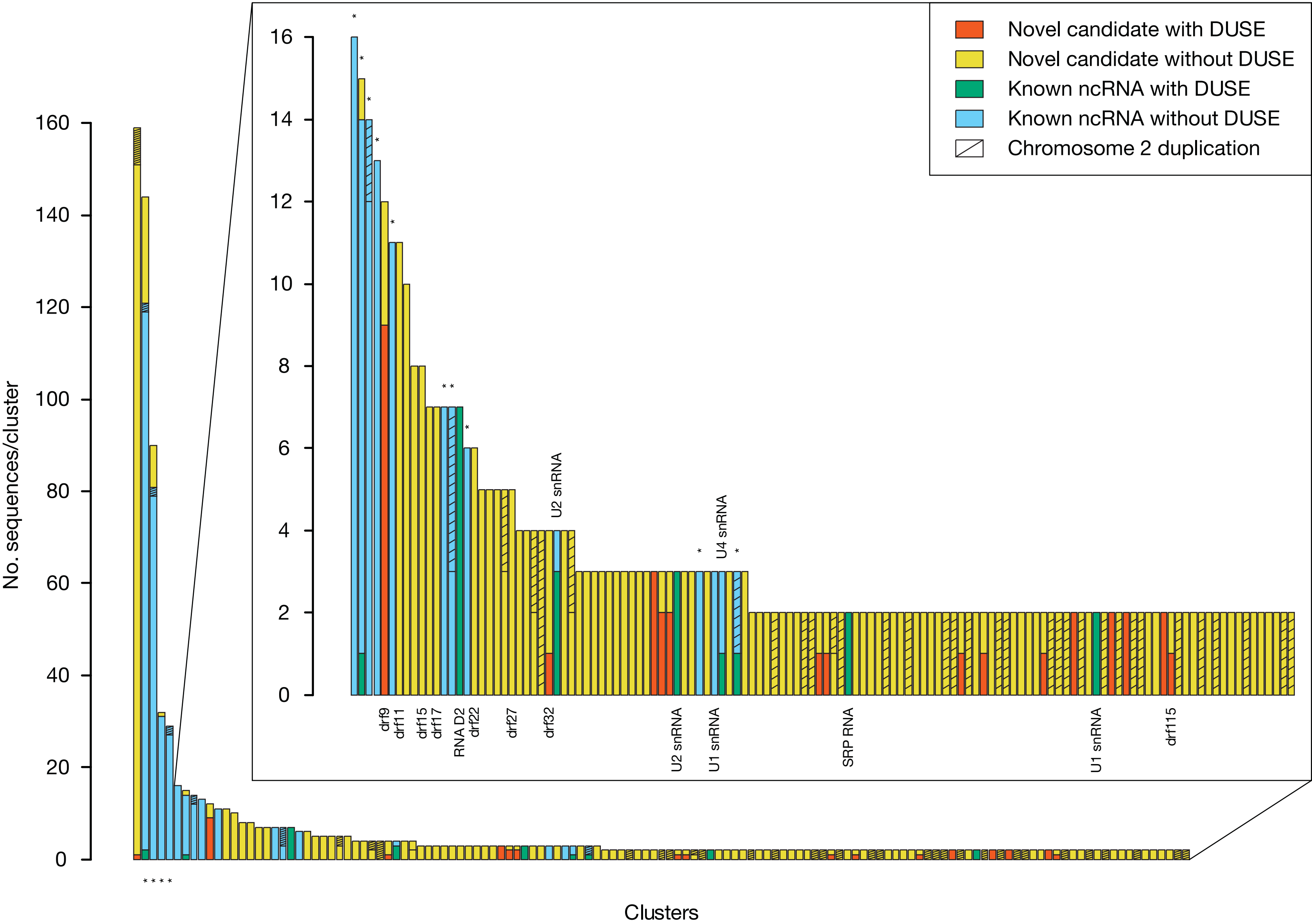
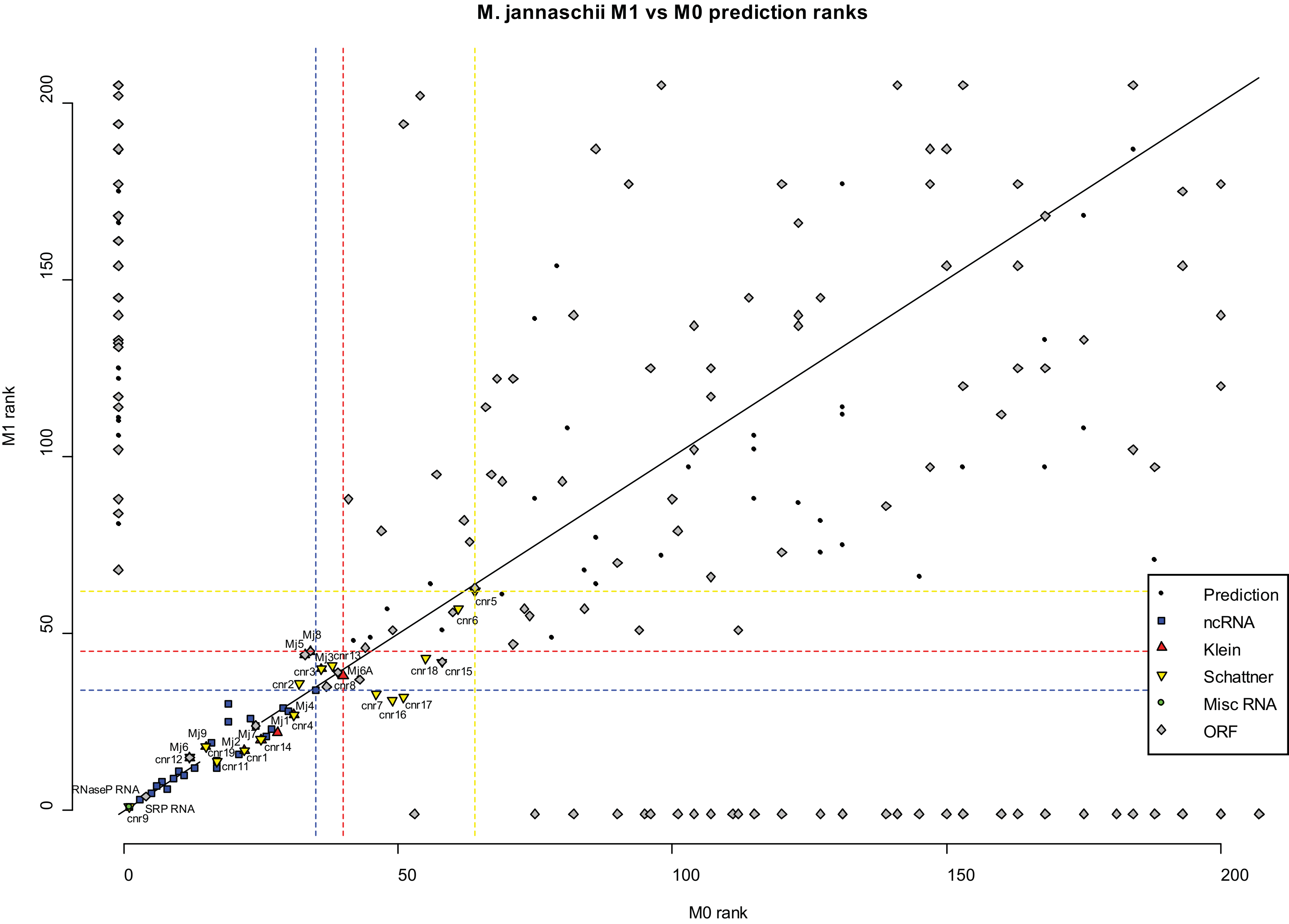
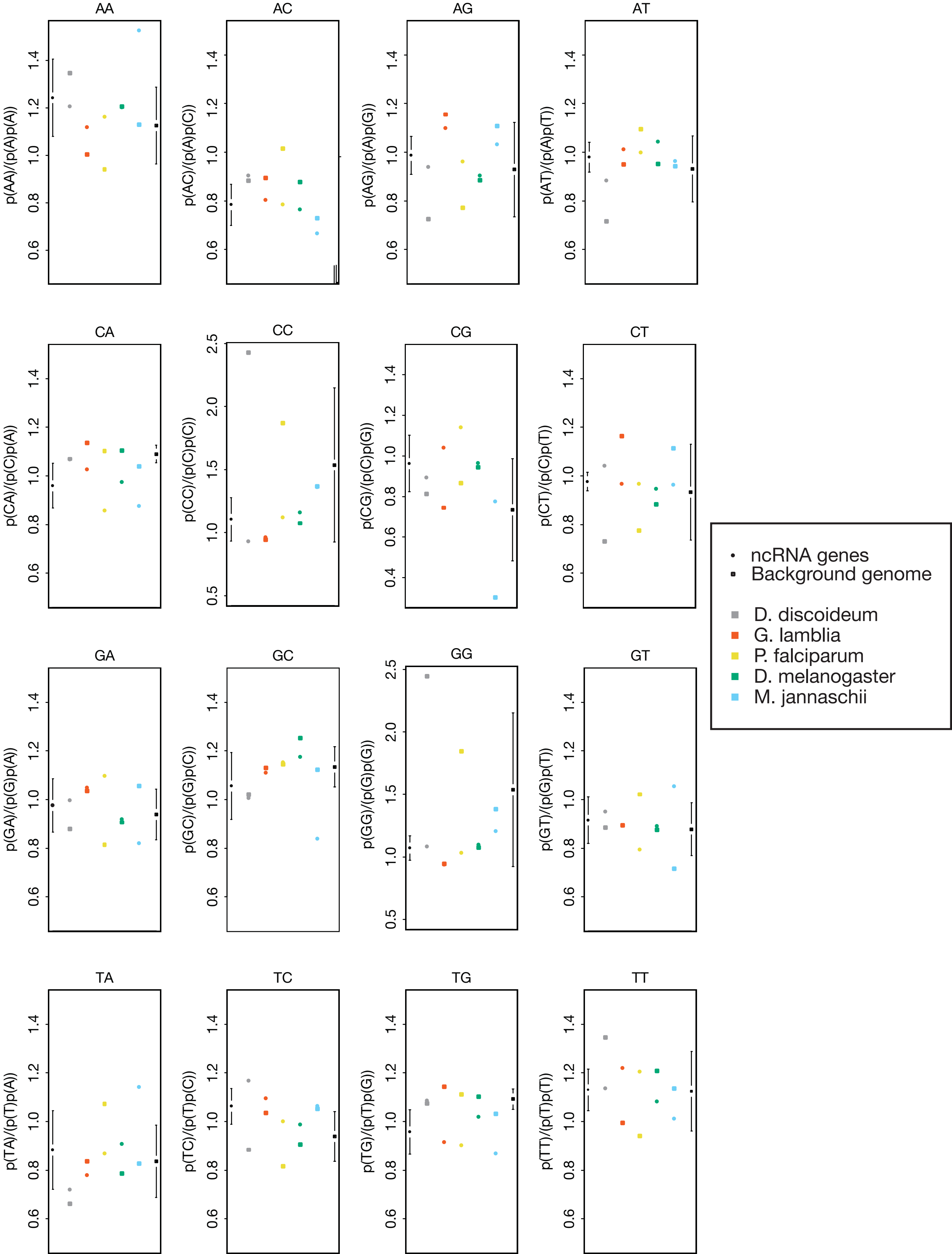
Supplemental Figure S1. Bar graph showing the size distribution and contents of clusters. Red and yellow colors indicate novel predictions with or without an upstream DUSE sequence, respectively. Green and blue bars represent predictions overlapping known or confidently predicted ncRNA genes with or without upstream DUSE sequence, respectively. Hatched bars indicate a sequence residing on the D. discoideum chromosome two duplication. Known or experimentally verified ncRNAs belonging to a cluster are indicated above or below bars. A star above or below a bar indicates presence of tRNAs within that cluster.



Supplemental Figure S2. Ranks of M. jannaschii M1 ncRNA predictions plotted against M0 ncRNA prediction. A solid black dot indicates a prediction. Gray diamonds indicates predictions that overlap known ORFs. Blue squares indicate predictions overlapping a known tRNA or rRNA gene. Red and yellow triangles represent predictions that overlap sequences reported in the works by Klein et al. and Schattner (Klein et al. 2002; Schattner 2002), respectively. Green circles represent SRP RNA and RNase P RNA. Identities of some predictions are indicated. The yellow and red dotted horizontal and vertical lines mark the ranks where all predictions from Schattner and Klein et al. are included. The blue dotted line indicates the ranks where all tRNA and rRNA genes are successfully predicted. A rank of -1 indicates that the region was not predicted by the corresponding model.



Supplemental Figure S3. Observed/expected dinucleotide ratio for ncRNAs and background genome for D. discoideum (Dd), G. lamblia (Gl), P. falcuiparum (Pf), D. melanogaster (Dm) and M. jannaschii (Mj). Filled circles represent observed/expected dinucleotide ratios for ncRNA genes and filled squares for background genome. Mean and standard deviation is indicated. Background genomes were constructed by masking annotated ncRNAs and exons except for M. jannaschii where only ncRNAs were masked. Sequence data and annotations were downloaded from the Generic Model Organism Database (http://gmod.mbl.edu), PlasmoDB (http://www.plasmodb.org), FlyBase (http://www.flybase.org) for Gl, Pf and Dm, respectively. Data for Mj was obtained from GenBank (http://www.ncbi.nlm.nih.gov).



Supplemental Data S1. Strand asymmetry

Interestingly, mononucleotide composition of the background genome is very nearly strand-symmetric, i.e. f(A) = f(T) = 0.43 and f(G) = f(C) = 0.069 within a strand (Sueoka's Parity Rule 2 (PR2) (Sueoka 1995)) while ncRNA copy strands (the DNA strand that corresponds to the RNA) are about 14% richer than their template strands in G. This suggests that the compositional contrast approach could in principle not only discriminate ncRNA gene-containing regions but also the template from the copy strand. Like mononucleotides, conditional dinucleotides exhibit substantial strand-asymmetry in the target but not the background, again suggesting a potential for strand-detection via contrast methods. However, when searching the genome using scores that were derived in a strand-dependent manner, the predictions of the coding strand did not turn out to be reliable.

Analyses of background and target ncRNA gene data. Perl script for calculating frequencies is available as Supplemental Data S2.

Observed-expected ratios of the target

	A	C	G	T
A	1.1588	0.7075	1.1121	0.9881
C	0.9748	1.0545	0.9614	1.0159
G	1.0004	0.9572	1.0296	1.0045
T	0.9170	1.2607	0.8330	1.0273

The G-test statistic G for the target = 542.0854 with 9 degrees of freedom.
The probability that sites are independent in the target is 0.0000

Observed-expected ratios of the background

	A	C	G	T
A	1.3476	0.8832	0.7252	0.7153
C	1.0695	2.4276	0.8127	0.7312
G	0.8799	1.0208	2.4461	0.8849
T	0.6617	0.8838	1.0733	1.3451

The G-test statistic G for the background = 1217561.4755 with 9 degrees of freedom.
The probability that sites are independent in the target is 0.0000

Conditional dinucleotide ratios of target over background

	A	C	G	T
A	0.4646	2.5359	6.0323	0.8858
C	0.4925	1.3752	4.6532	0.8910
G	0.6143	2.9687	1.6556	0.7280
T	0.7488	4.5161	3.0527	0.4898

Supplemental Data S3. Search for ncRNA genes in *M. jannaschii*

We compared our search method to the works of Klein *et al.* and Schattner (Klein et al. 2002; Schattner 2002) by searching for ncRNA genes in the AT-rich genome of *M. jannaschii*. Sequence data and annotations were downloaded from GenBank. We used the set of 43 tRNA and rRNA genes as target set and the genome with these ncRNA genes masked as background for calculating the scores. We subsequently searched the genome with both the M0 and the M1 model. The predictions are presented in Supplemental Fig. S2 where the ranks of the predictions for the different models are plotted against each other. Using both M0 and M1 we detect all of the tRNA and rRNA genes as well as all the five predicted ncRNAs that Klein *et al.* verified experimentally (some predictions overlap more than one gene). At this rank, the SRP RNA and RNase P RNA reported by Schattner along with 14 of his 19 predictions are also detected. All of the Schattner predictions are included along with 19 and 17 extra predictions at rank 64 and 62 for M0 and M1, respectively. It is interesting to note that the two predictions made by Klein *et al.* that overlap ORFs are the ones that score worst with the M1-model. Specifically, out of 207 and 205 predictions for M0 and M1, 59 predictions do not overlap neither annotated ORFs nor known ncRNAs or predictions by Klein *et al.* or Schattner. Of these, the M1 prediction has a lower rank in 46 cases. Hence, we readily reproduce the results of Klein *et al.* and Schattner using our approach.

Supplemental Table S1. Genomic coordinates and neighboring non-coding RNA genes of experimentally tested candidates. Strand indications are relative to the official v2.5 *D. discoideum* genome release.

	M1 prediction				Upstream			Downstream		
Name	Chr	Start	End	Strand	Gene	Strand	Distance	Gene	Strand	Distance
drd38	4	2475145	2475234	→						
drd3	1	1893172	1893369	→						
drf115_1	5	693177	693295	→				drf115_2		83
drf115_2	5	693378	693520		drf115_1	→	83			
drf17_1	1	1755846	1755987	←						
drf17_2	1	1756394	1756468	←						
drf17_3	1	4776055	4776158							
drf17_4	2	4773544	4773880							
drf17_5	5	849074	849617	→				drf17_6	←	1
drf17_6	5	849618	849833	←	drf17_5	→	1	drf17_7	←	482
drf17_7	5	850315	850389	←	drf17_6	←	482			
drf17_8	5	1065031	1065390							
drf27_1	2	87872	88021							
drf27_2	2	93620	93707							
drf27_3	2	291848	291997							
drf27_4	2	328690	328825							
drf27_5	4	3250352	3250527							
drf22_1	1	1651963	1652141	→						
drf22_2	3	774812	774995	→						
drf22_3	5	2268228	2268529	←						
drf22_4	5	2632109	2632192					drf22_5		1420
drf22_5	5	2633612	2633696		drf22_4		1420			
drf22_6	6	1904491	1904786	←						
drf9_1	2	4428918	4429023							
drf9_3	2	6866495	6866879							
drf9_4	3	1098356	1098859	←				ClassI Predicted	←	1534
drf9_5	3	1101912	1102184	→	ClassI Predicted	←	265			
drf9_6	4	897136	897398	→						
drf9_7	4	957509	957773		DdR-21	←	632	ClassI Predicted	←	471
drf9_8	4	960593	960976	→	DdR-22	←	255	ClassI Predicted	→	616
drf9_9	4	1662995	1663237	→	DdR-23C	←	249	DdR-24A	→	647
drf9_10	4	1667271	1667538	→	DdR-23B	←	270	ClassI Predicted	→	336
drf9_11	4	2792499	2792766	→	ClassI Predicted	←	261	drf9_12	←	399
drf9_12	4	2793165	2793326	→	drf9_11	→	399	ClassI Predicted	→	203
drf9_13	5	673919	674070	←						
drf15_1	1	1789139	1789262	→						
drf15_2	1	2135534	2135640							
drf15_3	2	89370	89506							
drf15_4	2	95426	95562							
drf15_5	2	326232	326368							
drf15_6	2	6241777	6241903							
drf15_7	3	2833138	2833245							
drf15_8	5	312466	312596							
drf32_1	1	2014925	2015070							
drf32_2	1	2858071	2858381							
drf32_3	2	8182027	8182378							
drf32_4	6	718936	719166							
drf11_1	2	294846	294985							
drf11_2	2	316677	317401							
drf11_3	2	323800	323886							
drf11_4	3	2830876	2831587							
drf11_5	3	2863427	2863537							
drf11_6	3	3665682	3665775							
drf11_7	3	4921473	4921646							
drf11_8	4	4276362	4276676							
drf11_9	5	2611116	2611334							
drf11_10	5	3909598	3909689							
drf11_11	5	4281170	4281589							

Supplemental Table S2. Oligonucleotides used for Northern blots, 5'RACE, and 3'RACE. Column "RNA": The RNA or group of RNA for which the oligos were used. 1) indicates hybridization signal > 500 nucleotides. 2 and 3 depict oligonucleotides also used for 3' and 5'RACE, respectively. For further details, see Material and Methods.

Name	Sequence (5'-3')	RNA	Hybridization
189	GGATTTGAAGTCCACACGCATC	drd38	+
206	GATGCGTGTGGACTTCAAATCC	drd38	-
190	GAACCAACTGTATGGATTTCTCC	drf115	+
284	GGAGAAATCCATACAGTTGGTTC	drf115	-
191	AGTCTCCCACCTATCCTACACAC	drd3	+
285	GTGTGTAGGATAGGTGGGAGACT	drd3	-
223	TACCCACCCATCCAACCCCAT	drf17	+
224	ATGGGGTTGGATGGGTGGGTA	drf17	-
227	AAATCCGGCTATTCGTCTTACAG	drf27	1) -
228	CTGTAAGACGAATAGCCGGATTT	drf27	1) -
229	CGACCGACCACTCCTCATTTAC	drf22	+
230	GTAAATGAGGAGTGGTCGGTCG	drf22	-
193	TTGTATTCTGGAGTCTGAATAGGT	drf9	+
207	ACCTATTCAGACTCCAGAATACAA	drf9	-
231[2)]	ACCTATTCAGACACCAGAATACAA	drf9	-
232[3)]	TTGTATTCTGGTGTCTGAATAGGT	drf9	+
233	CCCTTCCCATGGACTTCT	drf15	+
234	AGAAGTCCATGGGAAGGG	drf15	-
235	GGGAATCTATCCTGGTCAAATA	drf32	-
236	TATTTGACCAGGATAGATTCCC	drf32	-
239	TATACCCTGCAAGACATCCATTG	drf11	-
240	CAATGGATGTCTTGCAGGGTATA	drf11	-

Supplemental References

Klein, R.J., Z. Misulovin, and S.R. Eddy. 2002. Noncoding RNA genes identified in AT-rich hyperthermophiles. *PNAS* **99:** 7542-7547.

Schattner, P. 2002. Searching for RNA genes using base-composition statistics. *Nucl. Acids Res.* **30:** 2076-2082.

Sueoka, N. 1995. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *J Mol Evol* **40:** 318-325.