Supporting Information

Ivanov et al. 10.1073/pnas.1009269107

Table S1. Oligonucleotides used in this study

Number	Name	Sequence
1	AUU/S	GTAGGCCACCATTGCCG
	AUU/A	GATCCGGCAATGGTGGCCTACTGCA
2	UUG/S	GTAGGCCACCTTGGCCG
	UUG/A	GATCCGGCCAAGGTGGCCTACTGCA
3	CUG/S	GTAGGCCACCCTGGCCG
	CUG/A	GATCCGGCCAGGGTGGCCTACTGCA
4	AUC/S	GTAGGCCACCATCGCCG
	AUC/A	GATCCGGCGATGGTGGCCTACTGCA
5	AUA/S	GTAGGCCACCATAGCCG
	AUA/A	GATCCGGCTATGGTGGCCTACTGCA
6	AAG/S	GTAGGCCACCAAGGCCG
	AAG/A	GATCCGGCCTTGGTGGCCTACTGCA
7	AGG/S	GTAGGCCACCAGGGCCG
	AGG/A	GATCCGGCCCTGGTGGCCTACTGCA
8	ACG/S	GTAGGCCACCACGGCCG
	ACG/A	GATCCGGCCGTGGTGGCCTACTGCA
9	GUG/S	GTAGGCCACCGTGGCCG
	GUG/A	GATCCGGCCACGGTGGCCTACTGCA
10	GG/S	GTAGGCCGCCATGGCCG
	GG/A	GATCCGGCCATGGCGGCCTACTGCA
11	CG/S	GTAGGCCCCCATGGCCG
	CG/A	GATCCGGCCATGGGGGCCTACTGCA
12	UG/S	GTAGGCCTCCATGGCCG
	UG/A	GATCCGGCCATGGAGGCCTACTGCA
13	AA/S	GTAGGCCACCATGACCG
	AA/A	GATCCGGTCATGGTGGCCTACTGCA
14	AC/S	GTAGGCCACCATGCCCG
	AC/A	GATCCGGGCATGGTGGCCTACTGCA
15	AU/S	GTAGGCCACCATGTCCG
	AU/A	GATCCGGACATGGTGGCCTACTGCA
16	GA/S	GTAGGCCGCCATGACCG
	GA/A	GATCCGGTCATGGCGGCCTACTGCA
17	GC/S	GTAGGCCGCCATGCCCG
	GC/A	GATCCGGGCATGGCGGCCTACTGCA
18	GU/S	GTAGGCCGCCATGTCCG
	GU/A	GATCCGGACATGGCGGCCTACTGCA
19	CA/S	GTAGGCCCCCATGACCG
	CA/A	GATCCGGTCATGGGGGCCTACTGCA
20	CC/S	GTAGGCCCCCATGCCCG
	CC/A	GATCCGGGCATGGGGGCCTACTGCA
21	CU/S	GTAGGCCCCCATGTCCG
	CU/A	GATCCGGACATGGGGGCCTACTGCA
22	UA/S	GTAGGCCTCCATGACCG
	UA/A	GATCCGGTCATGGAGGCCTACTGCA
23	UC/S	GTAGGCCTCCATGCCCG
	UC/A	GATCCGGGCATGGAGGCCTACTGCA
24	UU/S	GTAGGCCTCCATGTCCG
	UU/A	GATCCGGACATGGAGGCCTACTGCA
25	AG/S	GTAGGCCACCATGGCCG
	AG/A	GATCCGGCCATGGTGGCCTACTGCA
26	EIF1 S	GTAGTATCGTATGTCCG
	EIF1 AS	GATCCGGACATACGATACTACTGCA
27	Neg Co S	GTAGGCCACCACAGCCG
	Neg Co AS	GATCCGGCTGTGGTGGCCTACTGCA
28	pSV40-Renilla S	GATCCTAATAGCACGTGTAATT
	pSV40-Renilla AS	CTAGAATTACACGTGCTATTAG
29	EIF1 WT S	CGCGCTAGCTAGTATCGTATGTCCGCTATCCAGAACCTCCAC
30	EIF1 good S	CGCGCTAGCTAGGCCACCATGGCCGCTATCCAGAACCTCCAC
31	EIF1 good* S	CGCGCTAGCTAGGCCACCATGTCCGCTATCCAGAACCTCCAC
32	EIF1 AS1	CGCTCTAGACACTTAAAACCCATGAACCTTCAGCTG

Fig. S1. Nucleotide sequences used for generating logograms: alignment of 18 nucleotides surrounding the AUG start codon of eIF1 from (*A*) metazoan species, (*B*) fungal species, (*C*) plant sequences excluding "eIF1 #2", and (*D*) plant sequences of "eIF1 #2." Species names are on the left. Nucleotide sequences from the corresponding eIF1 homolog are on the right. The initiation codon is highlighted in green. Nucleotides in positions −3/+4 are highlighted in red if they deviate from the Kozak consensus and in gray if they comply with it. A file with the full sequences is available upon request.

Fig. S1

Fig. S2. Consensus and eIF1 initiation contexts in seven different species. The consensus of initiation in different species was obtained from the Transterm database (1). The frequency of the occurrence of each nucleotide at each position is shown as a percentage of the total. The numbering, shown above, is relative to the "A" of the AUG codon. A consensus nucleotide occurring in more than 35% at a given position is highlighted in green. The nucleotides surrounding the start codon of the (main) eIF1 are highlighted as follows: magenta—those differing from the preferred consensus but occurring in more than 15% of other genes; red—those differing from the preferred consensus and occurring in less than 15% of other genes; yellow—those that coincide with the preferred consensus. (*A*) *Homo sapiens*, (*B*) *Drosophila melanogaster*, (*C*) *Neurospora crassa*, (*D*) *Toxoplasma gondii*, (*E*) *Oryza sativa*, (*F*) *Ustilago maydis*, and (*G*) *Saccharomyces cerevisiae*.

1. Jacobs GH, et al. (2009) Transterm: A database to aid the analysis of regulatory sequences in mRNAs. *Nucleic Acids Res* 37(Database issue):D72–D76.

Fig. S2

Fig. S3. Some plants have as many as three or even more paralogs of eIF1. Most of them are probably redundant, but one has a conservation pattern that suggests that it is not entirely redundant with the other(s). Because these eIF1 paralogs were not carefully classified before, we give them arbitrary numbers, 1–3, with each number corresponding to an orthologous group. For example, *Arabidopsis thaliana* paralogs 1–3 are currently annotated with gene IDs AT4G27130, AT5G54940, and AT1G54290. (*A*) A phylogenetic tree constructed with ClustalX from representative eIF1 protein sequences. The scale bar above represents the rate of divergence. The *Methanocaldococcus infernus* sequence is used to root the tree. Bootstrap values of key nodes are indicated. The branches linking paralogs of the plant "eIF1 #2" group are in red and those of the other paralogous groups are in green. As demonstrated by the branch lengths, plant "eIF1 #2" orthologs diverged from each other more quickly than did the other plant eIF1 paralogs and are therefore under more relaxed negative selection. For example, *A. thaliana* and *Raphanus sativus* paralogs #1 are 97% identical at the amino acid level and paralogs #3 are 98% identical, but paralogs #2 are only 90% identical, indicating that this paralog is diverging more quickly. This pattern of divergence is indicative of a nonredundant function for plant eIF1 paralog #2, which appears to be present only in some branches of flowering plants. (*B*) The protein sequences, in FASTA format, of the eIF1 homologs used to construct the tree shown in *A*. All plant eIF1 paralogs that do not belong to the group designated as #2 have their AUG initiation codon in a poor initiation context as shown in Fig. 1*C*. The 69 plant paralogs designated #2 for which sequence is available have an initiation consensus as shown in the logo in *C* and have an absolutely conserved +4 "G." Letter heights are proportional to the frequency of conservation. The consensus conservation is shown at the top of each logo; the position relative to the "A" of the AUG start codon is indicated below (crucial positions −3 and +4 are in red). Alignments of the sequences used to generate the logogram are shown in Fig. S1*D*.

Fig. S3