Supplementary materials to manuscript:
"Gene identification in novel eukaryotic genomes by self-training algorithm"

Table S1

A. thaliana

	number of exons	length aa	E-value	Domain name	Function
1	2	80	-	-	-
2	1	81	-	-	-
3	2	82	-	-	-
4	1	92	-	-	-
5	1	101	-	-	-
6	1	115	0.001	INT_IntI	(E2) integrases, site-specific tyrosine recombinases, DNA breaking-rejoining enzymes, N- and C-terminal domains.
7	1	219	2.0E-91	CAT	Chloramphenicol acetyltransferase.
8	1	219	2.0E-91	CAT	The same as #7.
9	1	385	0.003	Soj	Involved in chromosome partitioning and cell division
10	14	616	0.009	AsnB	Asparagine synthase (glutamine-hydrolyzing), amino acid transport and metabolism.

A. gambiae

	number of exons	length aa	E-value	Domain name	Function
1	2	73	2.0E-16	REX1	DNA repair REX1 is required for DNA repair in yeast
2	2	78	1.0E-15	COX6C	C oxidase subunit VIc cytochrome c oxidase, a 13 sub-unit complex, EC:1931 is the terminal oxidase in the mitochondrial electron transport chain.
3	2	79	-	-	-
4	3	79	-	-	-
5	2	83	4.0E-04	Complex1_LYR	Complex 1 protein (LYR family). This family of short proteins includes proteins from the NADH-ubiquinone oxidoreductase complex I. The family includes the B14 subunit from cow, and the B22 subunit from human.
6	2	88	-	-	-
7	2	89	5.0E-08	UPF0239	Protein family UPF0239
8	3	98	-	-	-
9	4	102	-	-	-
10	6	108	-	-	-
11	2	108	6.0E-05	CHCH	Conserved motif in the LOC118487, CHCH motif.
12	1	115	-	-	-
13	3	116	1.0E-13	Ribosomal_L44	Protein L44
14	1	128	-	-	-
15	1	135	-	-	-
16	2	141	-	-	-
17	3	150	-	-	-
18	2	159	-	-	-
19	2	164	-	-	-
20	3	178	1.0E-16	DUF423	With unknown function (DUF423). Potential integral membrane protein

21	3	201	-	-	-
22	3	202	0.001	MEA1	Enhanced antigen 1 (MEA1). Consists of several mammalian male enhanced antigen 1 (MEA1) proteins.
23	1	204	0.005	RimI	With unknown function
24	3	211	-	-	-
25	4	222	-	-	-
26	1	222	0.003	RimI	General function prediction only
27	1	227	4.0E-04	LITAF	Membrane-associated motif in LPS-induced tumor necrosis factor alpha factor (LITAF), also known as PIG7
28	2	229	0.002	Euk_Ferritin	Ferritin (Euk_Ferritin) domain. Primary iron storage proteins.
29	1	234	0.004	COG4934	Protease posttranslational modification, protein turnover, chaperones
30	1	264	-	-	-
31	3	290	-	-	-
32	4	324	2.0E-18	Transposase_1	This family includes the mariner transposase
33	1	326	-	-	-
34	3	330	6.0E-15	MRJP	Royal jelly protein, responsible for the high reproductive ability of the queen. Also the sequence-related yellow protein of drosophila which controls pigmentation of the adult cuticle and larval mouth parts.
35	2	355	-	-	-
36	3	364	1.0E-38	TAP42	The TOR signaling pathway activates a cell-growth program in response to nutrients TIP41 (pfam04176) interacts with TAP42 and negatively regulates the TOR signaling pathway.
37	1	365	-	-	-
38	4	373	-	-	-
39	2	375	-	-	-
40	1	390	4.0E-05	PgsA	Synthase lipid metabolism
41	2	416	-	-	-
42	1	454	-	-	-
43	1	466	-	-	-
44	1	470	4.0E-42	COG5062	Membrane protein. Function unknown
45	2	489	-	-	-
46	3	496	0.003	Amino_oxidase	Containing amine oxidoreductase. This family consists of various amine oxidases, including maze polyamine oxidase (PAO) and various flavin containing monoamine oxidases (MAO)
47	1	527	-	-	-
48	1	555	-	-	-
49	2	583	7.0E-19	EGL-9	Proline hydroxylase. Posttranslational modification, protein turnover, chaperones
50	1	698	-	-	-
51	3	727	-	-	-
52	1	733	0	Neurochondrin	This family contains several eukaryotic neurochondrin proteins.
53	4	741	0.001	FHA	Associated domain (FHA); found in eukaryotic and prokaryotic proteins
54	1	875	1.0E-143	Nup84_Nup100	Pore protein 84 / 107 Nup84p
55	8	1062	0.004	CA	Repeat domain; involved in Ca2+-mediated cell-cell adhesion; plays a role in cell fate, signalling, proliferation, differentiation, and migration

C. elegans

	number of	length aa	E-value	Domain name	Function

	exons				
1	2	90	-	-	-
2	2	91	-	-	-
3	2	92	-	-	-
4	2	94	-	-	-
5	3	97	-	-	-
6	2	99	-	-	-
7	1	101	-	-	-
8	2	102	0.009	SapB	Saposin (B) domain. Present in multiple copies in prosaposin and in pulmonary surfactant-associated protein B.
9	3	112	3.00E-04	SapB	The same as #8.
10	2	112	-	-	-
11	4	112	-	-	-
12	3	114	-	-	-
13	2	119	-	-	-
14	3	126	-	-	-
15	3	130	-	-	-
16	5	132	-	-	-
17	3	135	-	-	-
18	3	137	2.0E-07	TRS20	Subunit of TRAPP, an ER-Golgi tethering complex. Cell motility and secretion.
19	2	139	1.0E-04	DUF290	Transthyretin-like family. This family has weak similarity to transthyretin (formerly called prealbumin) which transports thyroid hormones.
20	2	140	-	-	-
21	2	141	-	-	-
22	2	142	0.003	DIM1	Mitosis protein DIM1.
23	2	143	-	-	-
24	2	145	-	-	-
25	2	152	-	-	-
26	4	154	-	-	-
27	4	157	-	-	-
28	3	160	-	-	-
29	6	172	-	-	-
30	4	180	-	-	-
31	3	184	-	-	-
32	5	185	-	-	-
33	4	222	-	-	-
34	4	224	-	-	-
35	6	229	-	-	-
36	4	238	-	-	-
37	5	265	-	-	-
38	5	273	-	-	-
39	4	302	-	-	-
40	4	308	-	-	-
41	7	340	-	-	-
42	1	402	2.0E-08	Transposase_11	Transposase DDE domain. Member of the DDE superfamily, which contains three carboxylate residues believed to be responsible for coordinating metal ions needed for catalysis.
43	1	402	2.0E-08	Transposase_11	The same as #42
44	9	578	0.008	RasGAP	GTPase-activator protein for Ras-like GTPases.
45	3	184	3.0E-94	-	TetC [Shigella flexneri], tetracycline resistance protein C in [Plasmid R100]

C. intestinalis

	number of exons	length aa	E-value	Domain name	Function
1	2	68	2.0E-07	GGL	involved in signal transduction via G-protein-coupled receptors
2	2	83	3.0E-14	zf-CSL	CSL zinc finger. The function is uncertain
3	2	100	-	-	-
4	1	102	5.0E-24	Chaperonin 10 Kd subunit	Involved in protein folding, ATP binding
5	1	112	-	-	-
6	1	131	-	-	-
7	3	131	2.0E-08	MAPEG family	Membrane associated proteins in eicosanoid and glutathione metabolism, catalyses the synthesis of PGE2 from PGH3
8	1	139	-	-	-
9	3	151	-	-	-
10	3	152	6.0E-13	MAPEG family	The same as #7
11	1	167	-	-	-
12	3	184	-	-	-
13	1	231	4.0E-61	PCMT	protein-L-isoaspartate (D-aspartate) O-methyltransferase activity
14	1	234	-	-	-
15	5	244	5.0E-09	L10	Ribosomal protein
16	3	249	2.0E-07	MIT	Microtubule interacting and trafficking molecule domain
17	1	253	-	-	-
18	1	286	0.009	RplO	Ribosomal protein L15 [Translation, ribosomal structure and biogenesis]
19	12	304	-	-	-
20	3	319	8.0E-60	Per1	A member of this family has been implemented in protein processing in the endoplasmic reticulum
21	5	357	-	-	-
22	9	397	-	-	-
23	1	407	3.0E-26	RNA_pol_I_A49	A49-like RNA polymerase I associated factor. Involved in transcription of ribosomal DNA
24	13	424	-	-	-
25	5	483	3.0E-06	PTB	Phosphotyrosine-binding (PTB) domain
26	1	517	3.0E-10	CAP_ED	Effector domain of the CAP family of transcription factors
27	2	764	2.0E-158	eIF3c_N	Eukaryotic translation initiation factor 3 subunit 8 N-terminus. The largest of the mammalian translation initiation factors

C. reinhardtii

	number of exons	length aa	E-value	Domain name	Function
1	3	56	3.0E-07	Ribosomal_S21e	Translation, ribosomal structure and biogenesis
2	2	64	1.0E-07	Nop10p	Nucleolar RNA-binding protein, Nop10p family. Essential for 18S rRNA production and rRNA pseudouridylation by the ribonucleoprotein particles containing H/ACA snoRNAs (H/ACA snoRNPs).
3	3	89	7.0E-15	CSL zinc finger	Unknown function
4	4	92	2.0E-08	Sec61beta family	Component of the Sec61/SecYEG protein secretory system
5	4	101	5.0E-04	Uncharacterized protein family	Unknown function

6	3	103	3.0E-17	ribosomal protein S21	Small ribosomal subunit
7	3	110	2.0E-18	SRP9	Signal recognition particle 9 kDa protein. Pausing of synthesis of ribosome associated nascent polypeptides that have been engaged by the targeting domain of SRP.
8	1	129	8.0E-07	ACN9 family	Localized to the mitochondrial intermembrane space may be a necessary general component of gluconeogenesis
9	3	130	1.0E-13	SCP2	SCP-2 sterol transfer family. Involved in binding sterols
10	5	134	7.0E-33	Ribosomal protein S9/S16	Ribosomal subunit
11	5	137	-	-	-
12	3	162	0.002	SAP	Predicted to be involved in chromosomal organization
13	4	236	7.0E-05	U1-like zinc finger	Zinc ion binding, nucleic acid binding

D. melanogaster

	number of exons	length aa	E-value	Domain name	Function
1	1	73	5.0E-17	REX1	DNA Repair. REX1 is required for DNA repair in yeast, and has homologues in other eukaryotes.
2	1	94	-	-	-
3	1	101	-	-	-
4	2	103	-	-	-
5	1	111	-	-	-
6	3	127	-	-	-
7	4	137	4.0E-05	Complex1_LYR	Complex 1 protein (LYR family). This family of short proteins includes proteins from the NADH-ubiquinone oxidoreductase complex I. The family includes the B14 subunit from cow, and the B22 subunit from human.
8	3	144	-	-	-
9	1	150	5.0E-09	Complex1_17_2kD	NADH:ubiquinone oxidoreductase 17.2 kD subunit. This family contains the 17.2 kD subunit of complex I and its homologues.