Supplementary tables to the paper: Computational identification of functional introns: high positional conservation of introns that harbor RNA genes

Michal Chorev, Liran Carmel

Species	Build	Source
Acyrthosiphon pisum	assembly2	AphidBase [1]
Aedes aegypti	AaegL1	Ensembl [2]
Ailuropoda melanoleuca	ailMel1	Ensembl [2]
Anolis carolinensis	anoCar1	Ensembl [2]
Anopheles gambiae	anoGam1	VectorBase [3]
Apis mellifera	apiMel3	Ensembl [2]
Aplysia californica	aplCal1	Ensembl [2]
Arabidopsis thaliana	TAIR10	Ensembl [2]
Aspergillus nidulans	ASM14920v1	Ensembl [2]
Bombyx mori	v2.0	SilkDB [4]
Bos Taurus	bosTau4	Ensembl [2]
Branchiostoma floridae	braFlo1	JGI [5]
Caenorhabditis brenneri	caePb2	UCSC [6]
Caenorhabditis briggsae	cb3	UCSC [6]
Caenorhabditis elegans	ce6	UCSC [6]
Caenorhabditis japonica	caeJap1	UCSC [6]
Caenorhabditis remanei	caeRem3	UCSC [6]
Callithrix jacchus	calJac3	Ensembl [2]

Canis familiaris	canFam2	Ensembl [2]
Cavia porcellus	cavPor3	Ensembl [2]
Choloepus hoffmanni	choHof1	Ensembl [2]
Ciona intestinalis	ci2	UCSC [6]
Ciona savignyi	CSAV2.0	Ensembl [2]
Danio rerio	danRer6	Ensembl [2]
Daphnia pulex	Dappu1	Ensembl [2]
Dasypus novemcinctus	dasNov2	Ensembl [2]
Dictyostelium discoideum	dictybase.01	Ensembl [2]
Dipodomys ordii	dipOrd1	Ensembl [2]
Drosophila ananassae	droAna3	FlyBase [7]
Drosophila erecta	droEre2	FlyBase [7]
Drosophila grimshawi	droGri2	FlyBase [7]
Drosophila melanogaster	dm3	UCSC [6]
Drosophila mojavensis	droMoj3	FlyBase [7]
Drosophila persimilis	droPer1	FlyBase [7]
Drosophila pseudoobscura	dp4	UCSC [6]
Drosophila sechellia	droSec1	FlyBase [7]
Drosophila simulans	droSim1	FlyBase [7]
Drosophila virilis	droVir3	FlyBase [7]
Drosophila yakuba	droYak2	FlyBase [7]
Echinops telfairi	TENREC	Ensembl [2]
Equus caballus	equCab2	Ensembl [2]
Erinaceus europaeus	eriEur1	Ensembl [2]

Felis catus	felCat3	Ensembl [2]
Fugu rubripes	fr2	UCSC [6]
Gallus gallus	galGal3	Ensembl [2]
Gasterosteus aculeatus	gasAcu1	Ensembl [2]
Giardia lamblia	1.1	Refseq [8]
Gorilla gorilla	gorGor3	Ensembl [2]
Homo sapiens	hg19,GRCh37	Ensembl [2]
Leishmania infantum	1.1	Ensembl [2]
Leishmania major	ASM272v2	Ensembl [2]
Linepithema humile	1.2	Fourmidable [9]
Loxodonta Africana	loxAfr3	Ensembl [2]
Macropus eugenii	Meug_1.0	Ensembl [2]
Microcebus murinus	micMur1	Ensembl [2]
Monodelphis domestica	monDom5	Ensembl [2]
Mus musculus	mm9	Ensembl [2]
Myotis lucifugus	myoLuc1	Ensembl [2]
Nematostella vectensis	Nemve1	Ensembl [2]
Neurospora crassa	ASM18292v1	Ensembl [2]
Ochotona princeps	OchPri2.0	Ensembl [2]
Ornithorhynchus anatinus	ornAna1	Ensembl [2]
Oryctolagus cuniculus	oryCun2.0	UCSC [6]
Oryza sativa	MSU6	Ensembl [2]
Oryzias latipes	oryLat2	Ensembl [2]
Otolemur garnettii	otoGar1	Ensembl [2]

Pan troglodytes	panTro2	Ensembl [2]
Petromyzon marinus	petMar1	Ensembl [2]
Phaeodactylum tricornutum	ASM15095v1	Ensembl [2]
Physcomitrella patens	ASM242v1	Ensembl [2]
Phytophthora infestans	ASM14294v1	Ensembl [2]
Plasmodium falciparum	ASM276v1	Ensembl [2]
Pogonomyrmex barbatus	1.2	Fourmidable [9]
Pongo abelii	ponAbe2	Ensembl [2]
Pristionchus pacificus	priPac1	Ensembl [2]
Procavia capensis	proCap1	Ensembl [2]
Pteropus vampyrus	pteVam1	Ensembl [2]
Puccinia graminis	ASM14992v1	Ensembl [2]
Puccinia triticina	ASM15152v1	Ensembl [2]
Rattus norvegicus	rn4	UCSC [6]
Rhesus macaque	rheMac2	Ensembl [2]
Saccharomyces cerevisiae	sacCer2	Ensembl [2]
Schistosoma mansoni	sma_v3.1	Ensembl [2]
Schizosaccharomyces pombe	1.1,ASM294v1	Refseq [8], Ensembl [2]
Sorex araneus	sorAra1	Ensembl [2]
Spermophilus tridecemlineatus	speTri1	Ensembl [2]
Strongylocentrotus purpuratus	strPur2	Ensembl [2]
Sus scrofa	susScr2	Ensembl [2]
Taeniopygia guttata	taeGut1	Ensembl [2]
Tetraodon nigroviridis	tetNig2	Ensembl [2]

Tribolium castaneum	Tcas3.0	BeetleBase [10]
Trichoplax adhaerens	ASM15027v1	Ensembl [2]
Tupaia belangeri	tupBel1	Ensembl [2]
Tursiops truncatus	turTru1	Ensembl [2]
Ustilago maydis	UM1	Ensembl [2]
Vicugna pacos	vicPac1	Ensembl [2]
Vitis vinifera	IGGP_12x	Ensembl [2]
Xenopus tropicalis	xenTro2	Ensembl [2]

Table S1. List of annotated genomes taken into our initial data set.

Species	# orthologous groups	Fraction
Physcomitrella patens	79	0.18
Plasmodium falciparum	108	0.24
Saccharomyces cerevisiae	121	0.27
Ustilago maydis	123	0.27
Schizosaccharomyces pombe	133	0.30
Phaeodactylum tricornutum	139	0.31
Neurospora crassa	140	0.31
Leishmania major	143	0.32
Arabidopsis thaliana	163	0.36
Oryza sativa	165	0.37
Vitis vinifera	167	0.37
Ciona savignyi	171	0.38
Dictyostelium discoideum	174	0.39
Phytophthora infestans	191	0.42
Caenorhabditis elegans	206	0.46
Schistosoma mansoni	213	0.47
Drosophila melanogaster	256	0.57
Xenopus tropicalis	336	0.75
Danio rerio	346	0.77
Ornithorhynchus anatinus	352	0.78
Gallus gallus	363	0.81
Anolis carolinensis	378	0.84

Mus musculus	394	0.88
Gasterosteus aculeatus	402	0.89
Monodelphis domestica	403	0.90
Pan troglodytes	409	0.91
Pongo abelii	427	0.95
Homo sapiens	450	1.00

Table S2. A list of the final 28 species used in the analysis. For each species, we provide the number of orthologous groups in which it has a representative gene, and the fraction it makes out of the 450 groups that were analyzed.

	miRNA-bearing	**miRNA-mixed**	**miRNA-lacking**	**total**
snoRNA-bearing	3	0	96	99
snoRNA-mixed	0	1	22	23
snoRNA-lacking	50	10	3,981	4,041
total	53	11	4,099	4,163

Table S3. Overlaps between miRNA/snoRNA-bearing/mixed/lacking unique patterns.

Feature	**Depicting**	**Description**
log-likelihood (LOGLIKE)	conservation	Given EREM's estimation of the evolutionary model parameters, this is the log-likelihood of observing the pattern, $\log l_p$.
log number of times observed	typicality	$\log n_p$. Combined with $\log l_p$, this gives, up to a multiplicative constant, the (log) ratio of the number of times we expect to observe that pattern in our data, divided by the actual observed number, $\log l_p - \log n_p = \log\left(\frac{l_p}{n_p}\right)$.
binomial test	typicality	p-value of the binomial test (with

		Bonferroni correction), measuring how likely it is to see the pattern n_p times, given its likelihood is l_p.
number of ones	conservation, antiquity	The number of 1's in the pattern.
fraction of ones (ONES_RATIO_KNOWN)	conservation, antiquity	The number of 1's divided by the total number of 1's and 0's in the pattern.
number of times observed	typicality	The number of occurrences of pattern p, n_p.
number of evolutionary events (gain weight 3, loss weight 1; SANKOFF_G3L1)	conservation	The minimum number of intron gain and loss events required to obtain the pattern, given that gains cost three times as much as losses (using the Sankoff algorithm).
number of evolutionary events (gain weight 1, loss weight 3; SANKOFF_G1L3)	conservation	The minimum number of intron gain and loss events required to obtain the pattern, given that losses cost three times as much as gains (using the Sankoff algorithm).
number of evolutionary events (gain weight 1, loss weight 1)	conservation	The minimum number of intron gain and loss events required to obtain the pattern, given that losses cost as

		much as gains (using the Fitch algorithm).
one in amphibians (IN_AMPHIBIAN)	conservation, antiquity	This feature is 1 if the pattern has a 1 in at least one amphibian (A. *carolinensis* or X. *tropicalis*), otherwise it is 0.
one in fish (IN_FISH)	conservation, antiquity	This feature is 1 if the pattern has a 1 in at least one fish (D. *rerio* or G. *aculeatus*), otherwise it is 0.
one in birds (IN_BIRD)	conservation, antiquity	This feature is 1 if the pattern has a 1 in G. *gallus* otherwise it is 0.
one in fungi (IN_FUNGI)	conservation, antiquity	This feature is 1 if the pattern has a 1 in at least one fungi (U. *maydis*, S. *pombe*, S. *cerevisiae*, or N. *crassa)*, otherwise it is 0.
one in plants (IN_PLANT)	conservation, antiquity	This feature is 1 if the pattern has a 1 in at least one plant (V. *vinifera*, A. *thaliana*, P. *patens*, or O. *sativa*), otherwise it is 0.
one in protists (IN_PROTIST)	conservation, antiquity	This feature is 1 if the pattern has a 1 in at least one protist (P. *falciparum*, D. *discoideum*, L. *major*, P. *tricornutum*, or P. *infestans*),

		otherwise it is 0.
one in U. *maydis*	conservation, antiquity	This feature is 1 if the pattern has a 1 in U. *maydis*, otherwise it is 0.
one in S. *pombe*	conservation, antiquity	This feature is 1 if the pattern has a 1 in S. *pombe*, otherwise it is 0.
one in S. *cerevisiae*	conservation, antiquity	This feature is 1 if the pattern has a 1 in S. *cerevisiae*, otherwise it is 0.
one in N. *crassa*	conservation, antiquity	This feature is 1 if the pattern has a 1 in N. *crassa*, otherwise it is 0.
one in V. *vinifera*	conservation, antiquity	This feature is 1 if the pattern has a 1 in V. *vinifera*, otherwise it is 0.
one in A. *thaliana*	conservation, antiquity	This feature is 1 if the pattern has a 1 in A. *thaliana*, otherwise it is 0.
one in P. *patens*	conservation, antiquity	This feature is 1 if the pattern has a 1 in P. *patens*, otherwise it is 0.
one in O. *sativa*	conservation, antiquity	This feature is 1 if the pattern has a 1 in O. *sativa*, otherwise it is 0.
one in P. *falciparum*	conservation, antiquity	This feature is 1 if the pattern has a 1 in P. *falciparum*, otherwise it is 0.
one in D. *discoideum*	conservation, antiquity	This feature is 1 if the pattern has a 1 in D. *discoideum*, otherwise it is 0.
one in L. *major*	conservation, antiquity	This feature is 1 if the pattern has a 1 in L. *major*, otherwise it is 0.

one in P. *tricornutum*	conservation, antiquity	This feature is 1 if the pattern has a 1 in P. *tricornutum*, otherwise it is 0.
one in P. *infestans*	conservation, antiquity	This feature is 1 if the pattern has a 1 in P. *infestans*, otherwise it is 0.
one in C. *elegans*	conservation, antiquity	This feature is 1 if the pattern has a 1 in C. *elegans*, otherwise it is 0.
one in C. savignyi	conservation, antiquity	This feature is 1 if the pattern has a 1 in C. *savignyi*, otherwise it is 0.
one in S. *mansoni*	conservation, antiquity	This feature is 1 if the pattern has a 1 in S. *mansoni*, otherwise it is 0.
one in G. *aculeatus*	conservation, antiquity	This feature is 1 if the pattern has a 1 in G. *aculeatus*, otherwise it is 0.
one in H. *sapiens*	conservation, antiquity	This feature is 1 if the pattern has a 1 in H. *sapiens*, otherwise it is 0.
one in A. *carolinensis*	conservation, antiquity	This feature is 1 if the pattern has a 1 in A. *carolinensis*, otherwise it is 0.
one in P. *troglodytes*	conservation, antiquity	This feature is 1 if the pattern has a 1 in P. *troglodytes*, otherwise it is 0.
one in P. *abelii*	conservation, antiquity	This feature is 1 if the pattern has a 1 in P. *abelii*, otherwise it is 0.
one in M. *domestica*	conservation, antiquity	This feature is 1 if the pattern has a 1 in M. *domestica*, otherwise it is 0.
one in M. *musculus*	conservation,	This feature is 1 if the pattern has a

	antiquity	1 in M. *musculus*, otherwise it is 0.
one in D. *melanogaster*	conservation, antiquity	This feature is 1 if the pattern has a 1 in D. *melanogaster*, otherwise it is 0.
one in D. *rerio*	conservation, antiquity	This feature is 1 if the pattern has a 1 in D. *rerio*, otherwise it is 0.
one in X. *tropicalis*	conservation, antiquity	This feature is 1 if the pattern has a 1 in X. *tropicalis*, otherwise it is 0.
intron density under LCA	conservation, antiquity	The last common ancestor (LCA) of all the intron-bearing species is assumed to be the species in which the intron was originated. LCA_LEAVES_RATIO is the ratio of intron-bearing (1's) to intron-lacking (0's) species, from among all the descendants of LCA.
taxonomic level of intron origin	antiquity	The number of ancestor nodes separating the LCA above from the tree root.
age of intron origin (LCA_AGE)	antiquity	The age of LCA above [MYA].
mean relative intron position	position	The mean distance of the exon-exon junction from the beginning of the

		coding sequence (CDS) divided by the CDS length.
median relative intron position (MED_REL_POSITION)	position	The median distance of the exon-exon junction from the beginning of the CDS divided by CDS length.
mean intron position	position	The mean distance of the exon-exon junction from the beginning of the CDS [nucleotides].
median intron position (MED_POSITION)	position	The median distance of the exon-exon junction from the beginning of the CDS [nucleotides].

Table S4. The initial set of 48 pattern-characterizing features. Abbreviated names are provided for the 13 features that were used in the final analysis. For each pattern, we note what aspect of the pattern it describes.

	Mean			median		
RNA gene	**-bearing pattern**	**-lacking pattern**	**p-value (t-test)**	**-bearing pattern**	**-lacking pattern**	**p-value (U-test)**
miRNA	690.99	400.37	$7.6 \cdot 10^{-4}$	454.9	184	$8.8 \cdot 10^{-9}$
snoRNA	840.18	392.52	$4.7 \cdot 10^{-16}$	744	184	$9.4 \cdot 10^{-22}$

Table S5. The results when LCA_AGE was calculated by EREM instead of by using the Dollo parsimony. Mean and median for both miRNA and snoRNA -bearing and -lacking unique patterns. P-values are Bonferroni-corrected.

References

1. Legeai, F., et al., *AphidBase: a centralized bioinformatic resource for annotation of the pea aphid genome.* Insect Mol Biol, 2010. **19 Suppl 2**: p. 5-12.
2. Flicek, P., et al., *Ensembl 2012.* Nucleic Acids Res, 2012. **40**(Database issue): p. D84-90.
3. Lawson, D., et al., *VectorBase: a data resource for invertebrate vector genomics.* Nucleic Acids Res, 2009. **37**(Database issue): p. D583-7.
4. Duan, J., et al., *SilkDB v2.0: a platform for silkworm (Bombyx mori) genome biology.* Nucleic Acids Res, 2010. **38**(Database issue): p. D453-6.
5. Grigoriev, I.V., et al., *The genome portal of the Department of Energy Joint Genome Institute.* Nucleic Acids Res, 2012. **40**(Database issue): p. D26-32.
6. Rhead, B., et al., *The UCSC Genome Browser database: update 2010.* Nucleic Acids Res, 2010. **38**(Database issue): p. D613-9.
7. McQuilton, P., S.E. St Pierre, and J. Thurmond, *FlyBase 101--the basics of navigating FlyBase.* Nucleic Acids Res, 2012. **40**(Database issue): p. D706-14.
8. Pruitt, K.D., T. Tatusova, and D.R. Maglott, *NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.* Nucleic Acids Res, 2007. **35**(Database issue): p. D61-5.
9. Wurm, Y., et al., *Fourmidable: a database for ant genomics.* BMC Genomics, 2009. **10**: p. 5.
10. Kim, H.S., et al., *BeetleBase in 2010: revisions to provide comprehensive genomic information for Tribolium castaneum.* Nucleic Acids Res, 2010. **38**(Database issue): p. D437-42.