Protein Cell 2011, 2(5): 395–409
DOI 10.1007/s13238-011-1052-z

Protein & Cell

RESEARCH ARTICLE

Spliceosomal genes in the *D. discoideum* genome: a comparison with those in *H. sapiens*, *D. melanogaster*, *A. thaliana* and *S. cerevisiae*

Bing Yu[1], Petra Fey[2], Karen E. Kestin-Pilcher[2], Alexei Fedorov[4], Ashwin Prakash[4], Rex L. Chisholm[2], Jane Y. Wu[3] ✉

[1] Department of Molecular and Clinical Genetics, Royal Prince Alfred Hospital and Sydney Medical School (Central), the University of Sydney, NSW 2006, Australia
[2] dictyBase, Center for Genetic Medicine, Northwestern University, Chicago, IL 60611, USA
[3] Department of Neurology and Lurie Comprehensive Cancer Center, Center for Genetic Medicine, Northwestern University Feinberg Medical School, Chicago, IL 60611, USA
[4] Department of Medicine and Program in Bioinformatics and Proteomics/Genomics, The University of Toledo, Toledo, OH 43614, USA
✉ Correspondence: jane-wu@northwestern.edu
Received May 11, 2011 Accepted May 20, 2011

ABSTRACT

Little is known about pre-mRNA splicing in *Dictyostelium discoideum* although its genome has been completely sequenced. Our analysis suggests that pre-mRNA splicing plays an important role in *D. discoideum* gene expression as two thirds of its genes contain at least one intron. Ongoing curation of the genome to date has revealed 40 genes in *D. discoideum* with clear evidence of alternative splicing, supporting the existence of alternative splicing in this unicellular organism. We identified 160 candidate U2-type spliceosomal proteins and related factors in *D. discoideum* based on 264 known human genes involved in splicing. Spliceosomal small ribonucleoproteins (snRNPs), PRP19 complex proteins and late-acting proteins are highly conserved in *D. discoideum* and throughout the metazoa. In non-snRNP and hnRNP families, *D. discoideum* orthologs are closer to those in *A. thaliana*, *D. melanogaster* and *H. sapiens* than to their counterparts in *S. cerevisiae*. Several splicing regulators, including SR proteins and CUG-binding proteins, were found in *D. discoideum*, but not in yeast. Our comprehensive catalog of spliceosomal proteins provides useful information for future studies of splicing in *D. discoideum* where the efficient genetic and biochemical manipulation will also further our general understanding of pre-mRNA splicing.

KEYWORDS pre-mRNA splicing, spliceosomal genes, *Dictyostelium discoideum*, comparative genomics, splicing regulators

INTRODUCTION

The amoeboid protozoan *Dictyostelium discoideum* is a eukaryotic model organism that has been extensively used in studying signal transduction, cell motility and cell differentiation. It occupies a unique phylogenetic position and belongs to the group of mycetozoans that branches out after plants but before metazoans and fungi (Baldauf et al., 2000). Little is known about the RNA processing machinery in *D. discoideum*.

Pre-mRNA splicing is the process that removes intervening sequences (introns) from the nascent pre-mRNA transcripts to form functional mRNAs. This process is a critical step in eukaryotic gene expression and occurs in the multi-component macromolecular machine named the spliceosome (e.g., Calarco. et al., 2011; Hoskins et al., 2011; Ramani. et al., 2011 and references within). This large RNA-protein complex contains, in addition to the pre-mRNA substrate, several uridine-rich small nuclear ribonucleoprotein (snRNP) particles as well as a number of associated proteins. To process the majority of introns (the major class, also called the U2-type introns), the spliceosome contains U1, U2, U4/6 and U5 snRNPs. The splicing of the minor class of

introns (also called the U12-type) occurs in the spliceosome containing U11 and U12 in addition to U4atac, U6atac and U5 snRNPs (for review, see (Patel and Steitz, 2003; Will Lührmann, 2005)). Biochemical and molecular studies have revealed major components of the splicing machinery, especially the U2-type spliceosome.

The completion of the *D. discoideum* genome (Eichinger et al., 2005) provides an opportunity for us to systematically examine pre-mRNA splicing and the splicing machinery in this model organism. We queried the *D. discoideum* genome available at dictyBase (http://dictybase.org; (Chisholm et al., 2006)) to determine the presence of introns in the coding sequences of the primary protein sequence set at dictyBase. The analysis revealed that among 13,527 predicted and known protein-coding genes in *D. discoideum*, 9232 (68%) contain at least one intron. This indicates that pre-mRNA splicing plays an important role in the expression of a majority of *D. discoideum* genes. Furthermore, in our comparison of genomic and expressed sequence tag (EST) sequences, we found that a number of *D. discoideum* genes undergo alternative pre-mRNA splicing, suggesting that alternative splicing regulation may play a role in the biology of this unicellular organism.

To identify genes encoding *D. discoideum* spliceosomal components, we searched dictyBase using sequences of spliceosomal proteins present in *Homo sapiens* (human). Our search criteria for *D. discoideum* orthologs included sequence similarity, reciprocal matches, the presence of the relevant domain(s), manual review and independent phylogenetic analysis. In general, we found that spliceosomal proteins and related factors in *D. discoideum* have higher similarity to those in the plant (*Arabidopsis thaliana*), fly (*Drosophila melanogaster*) and human (*Homo sapiens*) genomes than to their yeast (*Saccharomyces cerevisiae*) orthologs.

RESULTS AND DISCUSSION

D. discoideum, human, fly, plant and yeast genomes and their splicing features

D. discoideum has a genome size of 34 Mb, which is smaller than the human (2851 Mb), fly (180 Mb) and plant (157 Mb) genomes, but about 2.6 times the size of the yeast genome (13 Mb). The *D. discoideum* genome contains 13,527 predicted protein-coding genes, which is similar to those in fly (13,676) and plant (13,029), but significantly higher than those in yeast (5538) (Eichinger et al., 2005). We queried the *D. discoideum* genome sequence at dictyBase and found that in *D. discoideum*, 9210 (68%) contain at least one intron. It is known that 77% of fly genes (Crosby et al., 2007) and only 5% of yeast genes contain intron(s). The mean numbers of introns in spliced genes are 1.0, 1.9, 4.0 and 8.1 in yeast, *D. discoideum*, fly and human, respectively (Eichinger et al., 2005). We examined gene models and genomic sequences in comparison with ESTs and cDNA sequences. To date, this has led to the identification of 40 genes that have clear evidence of alternative splicing (Table 1). This is contrary to the previous belief that no regulated alternative splicing exists in any unicellular organism (Barbosa-Morais et al., 2006).

Based on published studies, we inspected protein sequences that have been reported as spliceosomal proteins or proteins with experimental evidence for their roles in pre-mRNA splicing. A collection of 264 human sequences for spliceosome associated proteins (Hartmuth et al., 2002; Zhou et al., 2002; Wu et al., 2004; Collins and Penny, 2005; Barbosa-Morais et al., 2006; Matlin and Moore, 2007; Bessonov et al., 2008) were retrieved from the RefSeq database and used to query the dictyBase database. Figure 1 shows a flow chart for our general search procedure. As a result, the vast majority of non-redundant homologs to human spliceosomal proteins and related factors (154) were identified in the *D. discoideum* genome. Furthermore, we identified several putative homologs by second-pass individual analyses (see METHODS). This increased the total number of putative spliceosomal proteins to 160. It demonstrates that 61% (160/264) of the human spliceosomal proteins have predicted orthologs in *D. discoideum*. The *D. discoideum* spliceosomal proteins and related factors are described below in several groups: the snRNP proteins, non-snRNP proteins, hnRNP and associated proteins, and alternative splicing regulators (Tables 2–5). "No hit" in Tables 2–5 indicates that the identified *D. discoideum* spliceosomal proteins did not hit their corresponding orthologs in fly, plant and yeast in the RefSeq database using our search criteria. In some cases, the fly, plant or yeast orthologs do exist, but are not identified using *D. discoideum* proteins because of the sequence divergence between *D. discoideum* and fly, plant or yeast.

Spliceosomal snRNP genes of *D. discoideum* are highly similar to their human orthologs

The snRNP proteins are further classified into Sm/Lsm core proteins, U1, U2, U5, U4/U6-specific proteins and tri-snRNP specific proteins (Table 2). Among snRNP proteins, all orthologs to 49 human proteins were identified in *D. discoideum*, plant, and fly genomes, but only 43 yeast orthologs were available (Table 2). Spliceosomal snRNP proteins in *D. discoideum* are highly conserved, and similar to those in higher eukaryotes. The Sm/LSm core proteins in the *D. discoideum* genome have almost one-to-one correspondences to their human counterparts. Such close relationship is illustrated by LSm6 (LSM6) and LSm7 (LSM7) in the phylogenetic tree (Fig. 2A).

When there are two or more closely related proteins in human, *D. discoideum* often has fewer, or just one ortholog. For example, searches with SmB/B' (SNRPB) or SmN (SNRPN) led to the same hit, DDB0233178 in *D. discoideum*. Similarly, only one gene with sequence similarity to

Table 1 Alternatively spliced genes in *D. discoideum*

Gene ID	dictyBase ID(s)	Gene name	Different proteins	References[a]
DDB_G0272560	DDB0185022, DDB0229383	capA	Yes	Bain et al., 1991
DDB_G0277501	DDB0185023, DDB0233889	capB	Yes	Escalante et al., 2003
DDB_G0277273	DDB0185226, DDB0231530	phg1b	Yes	Grant and Tsang, 1990
DDB_G0269160	DDB0191502, DDB0232009	nxnA	Yes	Greenwood and Tsang, 1991
DDB_G0267402	DDB0191157	h3a	No	
DDB_G0268044	DDB0349291, DDB0349293	DDB_G0268044	Yes	
DDB_G0268592	DDB0229901, DDB0304671, DDB0233887	dipA	Yes	
DDB_G0268708	DDB0349163, DDB0349164	DDB_G0268708	Yes	
DDB_G0269616	DDB0237972, DDB0237973	aplG	Yes	
DDB_G0274859	DDB0305016, DDB0231647	prdx4	Yes	
DDB_G0275533	DDB0308151, DDB0308152	DDB_G0275533	Yes	
DDB_G0277511	DDB0231604, DDB0231603	hpd	Yes	
DDB_G0277879	DDB0214911	ugpB	No	
DDB_G0277987	DDB0231342, DDB0347187	gxcA	Yes	p. c.[b]
DDB_G0278205	DDB0234174	eIF1a	No	
DDB_G0278645	DDB0238313, DDB0238314	DDB_G0278645	Yes	
DDB_G0279353	DDB0302646, DDB0302647	DDB_G0279353	Yes	
DDB_G0279427	DDB0232929	DDB_G0279427	No	
DDB_G0279557	DDB0305068, DDB0305069	DDB_G0279557	Yes	
DDB_G0279733	DDB0348652, DDB0348653	gxcDD	Yes	p. c.[b]
DDB_G0280495	DDB0231397, DDB0231400	fumH	Yes	
DDB_G0280765	DDB0233505	DDB_G0280765	No	
DDB_G0281325	DDB0305072	DDB_G0281325	No	
DDB_G0281587	DDB0238114, DDB0238121	yipf1	Yes	
DDB_G0283083	DDB0191416, DDB0229396	cbpD2	Yes	
DDB_G0284065	DDB0233724	DDB_G0284065	No	
DDB_G0284327	DDB0233634, DDB0233635	tmem184F	Yes	
DDB_G0284547	DDB0234051	DDB_G0284547	No	
DDB_G0284611	DDB0201626, DDB0304430	ripA	Yes	
DDB_G0286241	DDB0266659	aifD	No	
DDB_G0289173	DDB0191101	arfA	No	
DDB_G0289483	DDB0191103	ctxA	No	
DDB_G0289521	DDB0237692, DDB0237693	srsA	Yes	
DDB_G0289705	DDB0309126, DDB0309134	DDB_G0289705	Yes	
DDB_G0292188	DDB0234030, DDB0234031	DDB_G0292188	Yes	
DDB_G0293378	DDB0348900, DDB0348901	DDB_G0293378	Yes	
DDB_G0293526	DDB0201659	racC	No	
DDB_G0293770	DDB0237690, DDB0237691	DDB_G0293770	Yes	
DDB_G0295715	DDB0252885, DDB0252886	DDB_G0295715	Yes	
DDB_G0349057	DDB0349055, DDB0349059	DDB_G0349057	Yes	

[a] Reference is given if not first discovered in this analysis (dictyBase). So far, 40 genes with evidence for alternative splicing have been identified. As indicated in the third column, alternative splicing of 28 genes may produce different protein products. Other splice variants have two or more different transcripts as evidenced by mRNA or ESTs, but their effects on protein production are not evident and remain unclear.

[b] p. c., private communication to dictyBase (Dr. D. Veltman, Beatson Inst. For Cancer research, Glasgow, UK).

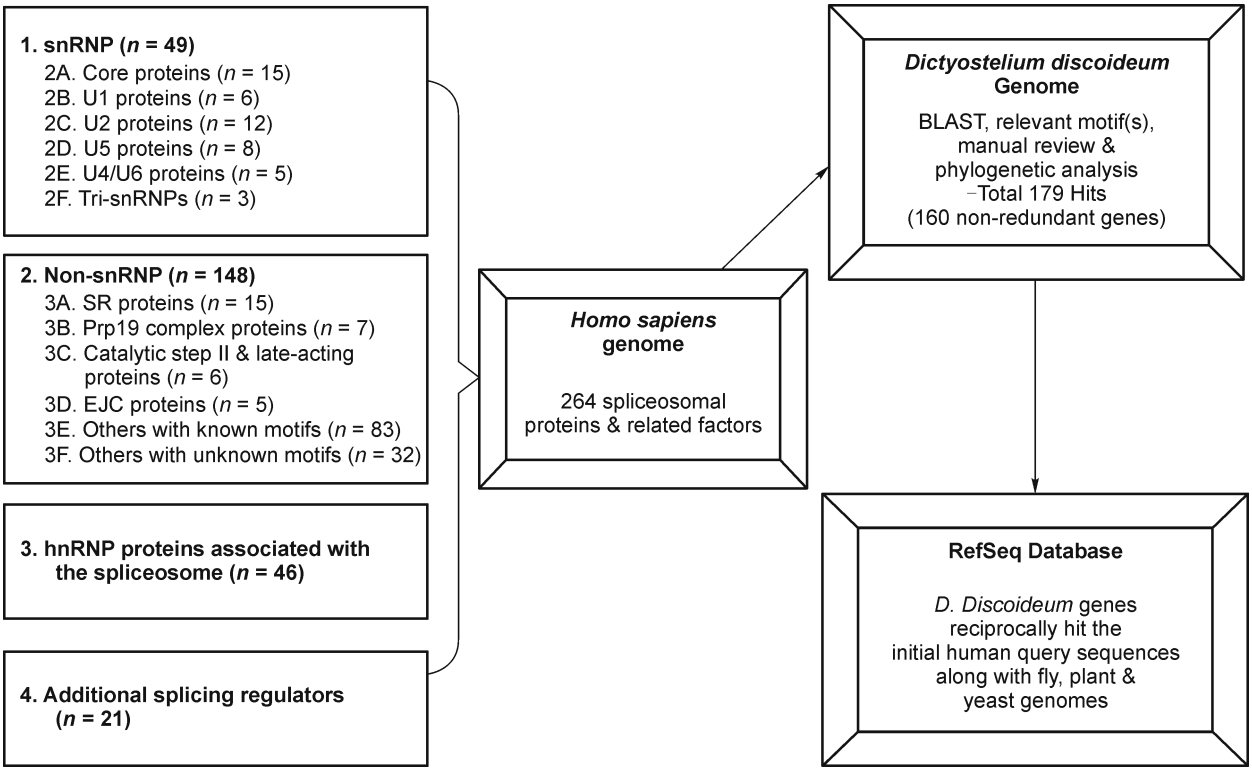

Figure 1. A schematic diagram of the search strategies used. The total number of 160 includes the initial 154 non-redundant hits and the additional 6 hits identified through the second-pass analysis.

Table 2 Spliceosomal snRNP proteins

Dictyostelium discoideum		*Human gene symbol*	*Homo sapiens*	*Drosophila melanogaster*	*Arabidopsis thaliana*	*Saccharomyces cerevisiae*
dictyBase ID[a]	Protein name[b]	HUGO[c]	RefSeq[a]	RefSeq[a]	RefSeq[a]	RefSeq[a]
2A. Sm/LSm core proteins						
DDB0233178-a	Sm B/B' (Smb1p)	SNRPB	NP_003082	*NP_476921*	*NP_199263*	*NP_010946*
DDB0233192	Sm D1 (Smd1p)	SNRPD1	NP_008869	NP_524774	NP_192193	NP_011588
DDB0233193	Sm D2 (Smd2p)	SNRPD2	NP_808210	NP_649645	NP_850477	NP_013377
DDB0233194	Sm D3 (Smd3p)	SNRPD3	NP_004166	NP_725106	NP_177757	NP_013248
DDB0233199	Sm F (Smx3p)	SNRPF	NP_003086	NP_523708	NP_194751	NP_015508
DDB0233197	Sm G (Smx2p)	SNRPG	NP_003087	NP_573139	NP_17997	NP_116636
DDB0233178-a	Sm N (Smb1p)	SNRPN	NP_003088	*NP_476921*	*NP_199263*	*NP_010946*
DDB0233376	LSm2 (Lsm2p)	LSM2	NP_067000	NP_648570	NP_563682	NP_009527
DDB0233377	LSm3 (Lsm3p)	LSM3	NP_055278	NP_732931	NP_177812	NP_013543
DDB0233378	LSm4 (Lsm4p)	LSM4	NP_036453	NP_001027236	NP_198124	NP_011037
DDB0233383	LSm5 (Lsm5p)	LSM5	NP_036454	NP_648022	NP_199698	NP_011073
DDB0233198	LSm6 (Lsm6p)	LSM6	NP_009011	NP_611528	*NP_181909*	NP_010666
DDB0233196	LSm7 (Lsm7p)	LSM7	NP_057283	NP_609807	NP_178480	NP_014252
DDB0233384	LSm8 (Lsm8p)	NAA38	NP_057284	NP_647660	NP_176747	NP_012556
DDB0302415[d]	SmE (Sme1p)	SNRPE	NP_003085	NP_609162	NP_567844	NP_014802
2B. U1 snRNP-specific proteins						
DDB0233134	U1-70K (Snp1p)	SNRP70	NP_003080	NP_477205	NP_190636	NP_012203
DDB0233514	CROP (Luc7p)	LUC7L3	NP_006098	NP_572337	NP_199954	NP_010196
DDB0233135-b	U1 A (na)	SNRPA	NP_004587	*NP_511045*	*NP_180585*	No hit

(*Continued*)

Dictyostelium discoideum		*Human gene symbol*	*Homo sapiens*	*Drosophila melanogaster*	*Arabidopsis thaliana*	*Saccharomyces cerevisiae*
dictyBase ID[a]	Protein name[b]	HUGO[c]	RefSeq[a]	RefSeq[a]	RefSeq[a]	RefSeq[a]
DDB0233515	U1 C (Yhc1p)	SNRPC	NP_003084	NP_650767	NP_567250	NP_013401
DDB0233540-c	Fnbp3 (Prp40p)	PRPF40A	NP_060362	*NP_722868*	NP_188601	*NP_012913*
DDB0233540-c	HYPC (Prp40p)	PRPF40B	NP_001026868	*NP_722868*	NP_175113	*NP_012913*
2C. U2 snRNP-specific proteins						
DDB0233176	U2 A' (Lea1p)	SNRPA1	NP_003081	NP_610315	NP_172447	NP_015111
DDB0233135-b	U2 B" (Msl1p)	SNRPB2	NP_003083	*NP_511045*	*NP_180585*	NP_012274
DDB0233180	SF3a120 (Prp21p)	SF3A1	NP_005868	NP_650583	NP_172917	NP_012332
DDB0233181	SF3a66 (Prp11p)	SF3A2	NP_009096	NP_648603	NP_565747	NP_010241
DDB0229922	SF3a60 (Prp9p)	SF3A3	NP_006793	NP_477114	NP_196234	NP_010254
DDB0233168	SF3b155 (Hsh155p)	SF3B1	NP_036565	NP_608534	NP_201232	NP_014015
DDB0233170	SF3b145 (Cus1p)	SF3B2	NP_006833	NP_608739	NP_193897	NP_013967
DDB0233171	SF3b130 (Rse1p)	SF3B3	NP_036558	NP_728546	NP_567015	NP_013663
DDB0233172	SF3b49 (Hsh49p)	SF3B4	NP_005841	NP_511058	NP_179441	NP_014964
DDB0233174	SF3b10 (na)	SF3B5	NP_112577	NP_652189	NP_849379	No hit
DDB0233775	SF3b14a (na)	SF3B14	NP_057131	NP_648037	NP_196780	No hit
DDB0233777	SF3b14b (Rds3p)	PHF5A	NP_116147	NP_609038	NP_563782	NP_015419
2D. U5 snRNP-specific proteins						
DDB0233128	U5-220 kDa (Prp8p)	PRPF8	NP_006436	NP_610735	NP_178124	NP_012035
DDB0233131	U5-200 kDa (Brr2p)	SNRNP200	NP_054733	NP_650472	NP_200922	NP_011099
DDB0233220	U5-116 kDa (Snu114p)	EFTUD2	NP_004238	NP_651605	NP_172112	NP_012748
DDB0233221	U5-102 kDa (Prp6p)	PRPF6	NP_036601	NP_649073	NP_192252	NP_009611
DDB0219950	U5-100 kDa (Prp28p)	DDX23	NP_004809	NP_609888	NP_180929	NP_010529
DDB0233518	U5-40 kDa (na)	SNRNP40	NP_004805	NP_608501	NP_181905	No hit
DDB0233510	U5-15 kDa (Dib1p)	TXNL4A	NP_006692	NP_608830	NP_196446	NP_015407
DDB0233538[d]	U5-52 kDa (na)	CD2BP2	NP_006101	NP_609404	NP_568211	No hit[d]
2E. U4/U6 snRNP-specific proteins						
DDB0233224	U4/U6-90 kDa (Prp3p)	PRPF3	NP_004689	NP_649156	NP_174127	NP_010761
DDB0233058	U4/U6-60 kDa (Prp4p)	PRPF4	NP_004688	NP_648990	NP_181681	NP_015504
DDB0233222	U4/U6-61 kDa (Prp31p)	PRPF31	NP_056444	NP_648756	NP_564754	NP_011605
DDB0233541	U4/U6-20 kDa (Cph1p)	PPIH	NP_006338	NP_610224	NP_001078301	NP_010439
DDB0233512	U4/U6-15.5 kDa (Snu13p)	NHP2L1	NP_001003796	NP_524714	NP_193969	NP_010888
2F. U4/U6.U5 tri-snRNP-specific proteins						
DDB0216406	Tri-snRNP 110 kDa (Snu66p)	SART1	NP_005137	NP_723700	NP_197180	NP_014953
DDB0220679	Tri-snRNP 65 kDa (Sad1p)	USP39	NP_006581	NP_573334	NP_193966	NP_116660
DDB0238807[d]	Tri-snRNP 27 kDa	SNRNP27	NP_006848	NP_611839	NP_568856	No hit

[a] Italic accession number refers to the sequence being used or hit two or more times. The same redundant hits are indicated by a code after dashed line "-". These codes are consistently used across Tables 2–5. "No hit" just indicates that no ortholog of that particular *D. disocideum* protein can be identified in fly, plant or yeast. It does not exclude the existence of the ortholog in fly or yeast.

[b] Conventional protein name with the yeast homolog is included in the parenthesis. na: no available yeast homolog.

[c] HUGO-gene symbol approved by Human Genome Organization gene nomenclature committee. Underlined gene symbol is not approved.

[d] Identified through the second-pass analysis.

Table 3 Non-snRNP spliceosomal proteins

Dictyostelium discoideum		*Human gene symbol*	*Homo sapiens*	*Drosophila melanogaster*	*Arabidopsis thaliana*	*Saccharomyces cerevisiae*
dictyBase ID[a]	Protein name[b]	HUGO[c]	RefSeq[a]	RefSeq[a]	RefSeq[a]	RefSeq[a]
3A. SR and SR-related proteins						
DDB0233327	9G8[e]	SRSF7	NP_001026854	NP_723226	*NP_190918*	No hit
DDB0233352-d DDB0233351-e	Tra2-beta[e]	SRSF10	NP_001177934	*NP_476764*	*NP_190991*	No hit
DDB0233352-d DDB0233351-e	Tra2-alpha	TRA2A	NP_037425	*NP_476764*	*NP_190991*	No hit
3B. PRP19 complex proteins						
DDB0233615-o	CDC5	CDC5L	NP_001244	NP_612033	NP_172448	NP_013940
DDB0233387-f	Prp5	DDX46	NP_055644	NP_573020	NP_187573	NP_009796
DDB0233530	Prp46	PLRG1	NP_002660	NP_572778	NP_566557	NP_015174
DDB0233583	fSap33	ISY1	NP_065752	NP_649768	NP_188509	NP_012584
DDB0233121	Prp19	PRPF19	NP_055317	NP_523783	NP_850206	NP_013064
DDB0233480	Crn	CRNKL1	NP_057736	NP_477118	NP_198992	NP_013218
DDB0233461	SYF1	XAB2	NP_064581	NP_610891	NP_198226	NP_010704
3C. Catalytic step II and late-acting proteins						
DDB0233399	Prp22	DHX8	NP_004932	*NP_610928*	NP_189288	*NP_010929*
DDB0233403	Prp43	DHX15	NP_001349	NP_610269	NP_191790	NP_011395
DDB0191460	Prp16	DHX38	NP_054722	NP_572947	NP_196805	NP_013012
DDB0233423	MCG9280	SLU7	NP_006416	NP_651659	NP_564859	NP_010373
DDB0233421	Prp17	CDC40	NP_056975	NP_651005	NP_172528	NP_010652
DDB0233422	Prp18	PRPF18	NP_003666	NP_650776	NP_563676	NP_011520
3D. Exon junction complex proteins						
DDB0233615-o	SRm160[e]	SRRM1	NP_005830	NP_648627	NP_180484	No hit
DDB0190682	UAP56	BAT1	NP_004631	*NP_723089*	*NP_191975*	*NP_010199*
DDB0233749	LDC2	RNPS1	NP_006702	NP_649903	*NP_565399*	No hit
DDB0233616	Y14	RBM8A	NP_005096	NP_610454	NP_564591	No hit
DDB0233620	Magoh	MAGOH	NP_002361	NP_476636	NP_171716	No hit
3E. Other spliceosomal proteins with known motifs						
DExD/H						
DDB0233447-g	DDX3	DDX3X	NP_001347	*NP_536783*	NP_567067	*NP_015206*
DDB0233447-g	DBY	DDX3Y	NP_004651	*NP_536783*	NP_181780	*NP_015206*
DDB0233431-h	p68[e]	DDX5	NP_004387	*NP_648062*	*NP_175911*	*NP_014287*
DDB0233431-h	p72[e]	DDX17	NP_006377	*NP_648062*	*NP_175911*	*NP_014287*
DDB0215338	Dbp5	DDX19B	NP_001014449	NP_001015151	NP_188610	NP_014689
DDB0233765-i	DICE1	INTS6	NP_036273	*NP_572253*	No hit	No hit
DDB0233765-i	FLJ41215	DDX26B	NP_872346	*NP_572253*	No hit	No hit
DDB0233387-f	DDXL	DDX39	NP_005795	*NP_723089*	*NP_191975*	*NP_010199*
DDB0233452	Abstrakt	DDX41	NP_057306	NP_524220	NP_199941	*NP_014287*
DDB0191511	NMP265	EIF4A3	NP_055555	NP_649788	NP_566469	NP_010304
DDB0233740	RHA	DHX9	NP_001348	NP_476641	NP_178223	NP_013523
DDB0233398	Prp2	DHX16	NP_003578	*NP_609946*	*NP_181077*	*NP_010929*
DDB0233419	DDX 35	DHX35	NP_068750	*NP_609946*	NP_174527	*NP_010929*
DDB0233453	fSAP118	SKIV2L2	NP_056175	NP_524929	NP_565338	*NP_012485*
DDB0233404	FLJ21972	DHX33	NP_064547	*NP_610928*	*NP_181077*	*NP_010929*
DDB0233432	SF3b125	DDX42	NP_031398	NP_648413	NP_566099	*NP_014287*

(*Continued*)

Dictyostelium discoideum		*Human gene symbol*	*Homo sapiens*	*Drosophila melanogaster*	*Arabidopsis thaliana*	*Saccharomyces cerevisiae*
dictyBase ID[a]	Protein name[b]	HUGO[c]	RefSeq[a]	RefSeq[a]	RefSeq[a]	RefSeq[a]
DDB0233454	SKI2W	SKIV2L	NP_008860	NP_524465	NP_190280	*NP_012485*
Cyclophilins						
DDB0191208	CyPL1	PPIL1	NP_057143	NP_523874	*NP_190046*	*NP_013633*
DDB0233543	CyP60	PPIL2	NP_055152	NP_611113	NP_201554	No hit
DDB0233550	CyPJ	PPIL3	NP_115861	NP_724939	NP_171696	No hit
DDB0233542	CyP64	PPWD1	NP_056157	NP_611935	*NP_190046*	NP_013317
DDB0233750	NY-CO-10	CWC27	NP_005860	NP_648508	NP_195032	No hit
WD40s						
DDB0185151	TEX1	THOC3	NP_115737	NP_649784	NP_200424	No hit
DDB0233537	fSAP57	SMU1	NP_060695	NP_650766	NP_177513	NP_014082
DDB0233739	fSAP35	THOC6	NP_077315	NP_648557	NP_190535	No hit
DDB0233741	MGC4238	WDR83	NP_115708	NP_609782	NP_193325	No hit
Cap binding proteins						
DDB0233457	CBP80	NCBP1	NP_002477	NP_726938	NP_565356	NP_013844
DDB0233456	CBP20	NCBP2	NP_031388	NP_524396	NP_199233	NP_015147
Polyadenylation machinery						
DDB0233359	PAB1	PABPC1	NP_002559	NP_476667	*NP_177322*	NP_011092
DDB0233678	PAB2	PABPN1	NP_004634	NP_476902	NP_201329	NP_014394
DDB0204541	CPSF 160	CPSF1	NP_037423	NP_995833	NP_199979	NP_010587
ZnF motif						
DDB0233743	ZNF183	RNF113A	NP_008909	NP_650865	No hit	NP_013427
DDB0233459	fSAP47	RBM22	NP_060517	NP_649440	NP_563788	NP_009621
DDB0233683	ZNF207	ZNF207	NP_001027464	NP_001097167	NP_564369	NP_012479
DDB0216436	FLJ31121	ZMAT2	NP_653324	NP_647881	NP_566257	NP_010185
Other motifs						
DDB0233354	U2 AF35	U2AF1	NP_001020374	NP_477208	No hit	No hit
DDB0233353	U2 AF65	U2AF2	NP_001012496	NP_476891	NP_176287	No hit
DDB0233439	fSAP164	AQR	NP_055506	NP_731647	NP_850297	*NP_013797*
DDB0220666	fSAP152	ACIN1	NP_055792	NP_609935	*NP_195678*	No hit
DDB0233680	SPF45[e]	RBM17	NP_116294	NP_001036426	NP_174336	No hit
DDB0252841	fSAP94	RBM25	NP_067062	NP_572242	NP_195370	No hit
DDB0233391	ZFM1[e]	SF1	NP_004621	*NP_524654*	NP_199943	NP_013217
DDB0252764	fSAPa	SR140	NP_001073884	NP_611535	No hit	No hit
DDB0233213	TAT-SF1	HTATSF1	NP_055315	NP_649313	NP_197130	NP_014113
DDB0233350	fSAP59	RBM39	NP_004893	NP_609095	*NP_565399*	No hit
DDB0191206	SKIP	SNW1	NP_036377	NP_511093	NP_177845	NP_009370
DDB0233560	Hpr1	THOC1	NP_005122	NP_649594	NP_568219	No hit
DDB0233565	fSAPb	CCDC40	NP_060420	NP_609877	NP_178208	NP_011794
DDB0229428	CRK7	CDK12	NP_057591	NP_649325	NP_192739	NP_012783
DDB0233563	TIP39	TFIP11	NP_001008697	NP_524725	NP_181762	No hit
DDB0216281	Prp4 kinase	PRPF4B	NP_003904	NP_612010	NP_189213	No hit
DDB0233564	CPSF5	NUDT21	NP_008937	NP_001036597	NP_194285	No hit
DDB0233566	fSAPc	C19orf29	NP_067054	NP_523422	NP_171887	No hit
DDB0233745	FLJ10374	CCDC94	NP_060544	NP_611092	NP_173156	NP_012828
DDB0233752	RRP6	EXOSC10	NP_001001998	NP_001097795	NP_198440	NP_014643

(*Continued*)

Dictyostelium discoideum		*Human gene symbol*	*Homo sapiens*	*Drosophila melanogaster*	*Arabidopsis thaliana*	*Saccharomyces cerevisiae*
dictyBase ID[a]	Protein name[b]	HUGO[c]	RefSeq[a]	RefSeq[a]	RefSeq[a]	RefSeq[a]
DDB0219931	SMC1	SMC1A	NP_006297	NP_651211	NP_191027	NP_116647
DDB0219933	CAPE	SMC2	NP_006435	NP_610995	NP_190330	NP_116687
DDB0233753	HUB1	UBL5	NP_077268	NP_610239	NP_199045	NP_014430
DDB0233546	Prp38	PRPF38B	NP_060531	NP_572872	NP_851101	No hit
DDB0233547	Prp39	PRPF39	NP_060392	NP_651634	NP_563700	NP_013667
DDB0233523	fSAP17	BUD31	NP_003901	NP_511117	NP_193843	NP_009990
DDB0302412	MRPS4	IMP3	NP_060755	NP_611224	No hit	NP_012018
DDB0233746	fSAP113	MOV10	NP_066014	NP_647816	No hit	*NP_013797*
DDB0233601[d]	SPF30[e]	SMNDC1	NP_005862	No hit	NP_178361	No hit
DDB0302416[d]	CGI-25	NOSIP	NP_057037	NP_573288	NP_564781	No hit
3F. Other spliceosomal proteins without known motifs						
DDB0233742	ASR2	SRRT	NP_056992	NP_610203	NP_565635	No hit
DDB0233558	THO2	THOC2	NP_001075019	NP_722763	NP_173871	NP_014260
DDB0233570	fSAP24	THOC7	NP_079351	NP_728489	NP_568339	No hit
DDB0233571	fSAP71	BUD13	NP_116114	NP_651272	NP_174470	NP_011341
DDB0233575	MFAP1	MFAP1	NP_005917	NP_647679	NP_197292	No hit
DDB0233767	FEM-2	PPM1F	NP_055449	*NP_609899*	NP_568786	NP_010278
DDB0233582	fSAP105	C21orf66	NP_037461	NP_649689	No hit	No hit
DDB0233585	CCAP2	CWC15	NP_057487	NP_652605	NP_566447	NP_010447
DDB0233592	DGSI	DGCR14	NP_073210	NP_572480	NP_187436	No hit
DDB0233598	NAP	CTNNBL1	NP_110517	NP_649847	NP_186920	No hit
DDB0233748	SPF27	BCAS2	NP_005863	NP_651596	NP_566599	No hit
DDB0233764-j	LUCA15	RBM5	NP_005769	NP_722689	*NP_190991*	No hit
DDB0233764-j	KIAA0122	RBM10	NP_005667	NP_608582	*NP_190991*	No hit
DDB0233755	Lupus La	SSB	NP_003133	NP_477014	NP_178106	NP_010232
DDB0233034	Kin17	KIN	NP_036443	NP_649212	NP_564690	NP_014720
DDB0233783	G patch containing 1	GPATCH1	NP_060495	NP_648669	NP_197699	No hit
DDB0233744	FLJ39430	CCDC12	NP_653317	NP_651757	NP_566245	No hit
DDB0216283	fSAP23	GTF2B	NP_001505	NP_476888	NP_187644	NP_015411
DDB0233593[d]	fSAP29	SYF2	NP_056299	NP_610617	NP_565396	No hit

[a] The same as in Table 2.

[b] Conventional protein name.

[c] HUGO-gene symbol approved by Human Genome Organization gene nomenclature committee. Underlined gene symbol is not approved.

[d] Identified through the second-pass analysis.

[e] The non-snRNP spliceosomal protein is also involved in splicing regulation.

both SNRPB and SNRPN has been identified in the *D. melanogaster*, *A. thaliana* and *S. cerevisiae* genomes. This relationship has been demonstrated in the phylogenetic analysis (Fig. 2B). These orthologs in the fly, plant, *D. discoideum* and yeast genomes are arranged in the expected evolutionary position.

There is only one gene (DDB0233135) identified in *D. discoideum* with significant sequence similarity to both human U1-specific protein A (U1A, SNRPA) and U2-specific protein B2 (U2B", SNRPB2). This finding is similar to what is found in the fly and plant. Interestingly, this *D. discoideum* ortholog only identifies U2B" (Msl1p), but not U1A (Mud1p) in yeast. It is possible that other U1A like genes exist in *D. discoideum*, but with more divergent sequences. Three proteins, SmE (SNRPE), the U5-specific 52 kDa (CD2BP2) and tri-snRNP 27 kDa (SNRNP27), were identified by

Table 4 hnRNP proteins associated with the spliceosome

Dictyostelium discoideum		*Human gene symbol*	*Homo sapiens*	*Drosophila melanogaster*	*Arabidopsis thaliana*	*Saccharomyces cerevisiae*
dictyBase ID[a]	Protein name[b]	HUGO[c]	RefSeq[a]	RefSeq[a]	RefSeq[a]	RefSeq[a]
DDB0233648	hnRNP L	HNRNPL	NP_001005335	NP_476731	*NP_175010*	No hit
DDB0214833-k	hnRNP R	HNRNPR	NP_005817	*NP_650913*	*NP_567192* *NP_179916*	No hit
DDB0214833-k	hnRNP Q	SYNCRIP	NP_006363	*NP_650913*	*NP_567192* *NP_179916*	No hit
DDB0233126	Bub3	BUB3	NP_001007794	NP_477381	NP_175413	NP_014669
DDB0305338-l *DDB0191168-s* *DDB0238264-t*	HSP70	HSPA1A	NP_005336	NP_524339 NP_731651 NP_524927 NP_788663	*NP_173055* *NP_187864* *NP_187555* *NP_195870* *NP_195869*	*NP_011029* *NP_009478* *NP_009396* *NP_013076*
DDB0305338-l *DDB0191168-s* *DDB0238264-t*	HSP71	HSPA8	NP_006588	NP_524356 NP_524063 NP_729940	*NP_173055* *NP_187864* *NP_187555* *NP_195870* *NP_195869*	*NP_011029* *NP_009478* *NP_009396* *NP_013076*
DDB0233663	GRP78	HSPA5	NP_005338	NP_727563	NP_851119 NP_198206	NP_012500
DDB0233747	RNPL	RBM3	NP_006734	No hit	NP_851195	No hit
DDB0233645-m	hnRNP I[e]	PTBP1	NP_002810	*NP_524703*	*NP_175010*	No hit
DDB0233645-m	nPTB	PTBP2	NP_067013	*NP_733460*	*NP_175010*	No hit
DDB0233645-m	PTBP3	ROD1	NP_005147	*NP_524703*	*NP_175010*	No hit
DDB0233651-n	HuR	ELAVL1	NP_001410	*NP_476936*	NP_188544	No hit
DDB0233651-n	HuB[e]	ELAVL2	NP_004423	*NP_476936*	*NP_173690*	No hit
DDB0233651-n	HuC	ELAVL3	NP_001411	*NP_476936*	*NP_173690*	No hit
DDB0233651-n	HuD	ELAVL4	NP_068771	*NP_476936*	NP_195137	No hit
DDB0232001	Ku70	XRCC6	NP_001460	NP_536773	NP_564012	NP_014011

[a] The same as in Table 2.

[b] Conventional protein name.

[c] HUGO-gene symbol approved by Human Genome Organization gene nomenclature committee.

[d] Identified through the second-pass analysis.

[e] The H complex protein is also involved in splicing regulation.

second-pass blast and domain analyses (see METHODS). The small (92 amino acids) SNRPE homolog (DDB0302415) was not present in *D. discoideum* gene predictions but was identified using the human SNRPE protein sequence and the tBlastn program. The putative *Dictyostelium* CD2BP2 homolog (DDB0233538) contains a > 50 polyasparagine stretch. Homopolymers, especially polyglutamine and polyasparagine stretches are abundant in *D. discoideum* (Eichinger et al., 2005). These proteins can be classified by customized blast searches (see METHODS). The possible SNRNP27 homolog (DDB0238807) in *D. discoideum*, which is present in fly but absent in yeast, was also identified using the tBlastn program. This revealed a gene where the original gene structure was incorrect. After curation at dictyBase, the predicted homolog aligns well with the human SNRNP27 at the C-terminus, where both proteins contain a DUF1777 domain whose function remains unclear. The *D. discoideum* protein is almost twice as long as its human counterpart (296 *versus* 155 amino acids). However, this difference in length occurs in the repetitive arginine-rich N-terminal sequences that both proteins share to a different degree.

Non-snRNP proteins associated with spliceosomal assembly and splicing

Non-snRNP proteins associated with spliceosomal assembly and pre-mRNA splicing are classified into several groups: SR and SR-related proteins, PRP19 complex proteins, catalytic step II and late-acting proteins, exon junction complex (EJC) proteins and other splicing factors. We searched the *D. discoideum* proteome using corresponding human proteins with the same criteria as described above. The *D. discoideum* non-snRNP proteins are more similar to *A. thaliana*, *D. melanogaster* and *H. sapiens* orthologs than are those of

Table 5 Additional proteins involved in alternative splicing regulation

Dictyostelium discoideum		*Human gene symbol*	*Homo sapiens*	*Drosophila melanogaster*	*Arabidopsis thaliana*	*Saccharomyces cerevisiae*
dictyBase ID[a]	Protein name[b]	HUGO[c]	RefSeq[a]	RefSeq[a]	RefSeq[a]	RefSeq[a]
DDB0233674-o *DDB0233675-p*	CUG-BP	CELF1	NP_001020767	*NP_788039* *NP_609559* *NP_723739*	*NP_171845* *NP_567249* *NP_973752*	No hit
DDB0233674-o *DDB0233675-p*	ETR3	CELF2	NP_001020247	*NP_788039* *NP_609559* *NP_723739*	*NP_171845* *NP_567249* *NP_973752*	No hit
DDB0233674-o *DDB0233675-p*	CAGH4	CELF3	NP_009116	*NP_788039* *NP_609559* *NP_723739*	*NP_171845* *NP_567249* *NP_973752*	No hit
DDB0233674-o *DDB0233675-p*	Bruno-like 4	CELF4	NP_064565	*NP_788039* *NP_609559* *NP_723739*	*NP_171845* *NP_567249* *NP_973752*	No hit
DDB0233674-o *DDB0233675-p*	Bruno-like 5	CELF5	NP_068757	*NP_788039* *NP_609559* *NP_723739*	*NP_171845* *NP_567249* *NP_973752*	No hit
DDB0233674-o *DDB0233675-p*	Bruno-like 6	CELF6	NP_443072	*NP_788039* *NP_609559* *NP_723739*	*NP_171845* *NP_567249* *NP_973752*	No hit
DDB0230135-q	SFRSK1	SRPK1	NP_003128	*NP_725459*	NP_197675	*NP_013943*
DDB0230135-q	SFRSK2	SRPK2	NP_872633	*NP_725459*	NP_566977	*NP_013943*
DDB0230105-r	CDC-like kinase 1	CLK1	NP_004062	*NP_001014680* *NP_001014681*	*NP_194205*	*NP_013081*
DDB0230105-r	CDC-like kinase 2	CLK2	NP_003984	*NP_001014680* *NP_001014681*	*NP_194205*	*NP_013081*
DDB0230105-r	CDC-like kinase 3	CLK3	NP_003983	*NP_001014680* *NP_001014681*	*NP_194205*	*NP_013081*
DDB0230105-r	CDC-like kinase 4	CLK4	NP_065717	*NP_001014680* *NP_001014681*	*NP_194205*	*NP_013081*

[a] The same as in Table 2.

[b] Conventional protein name.

[c] HUGO-gene symbol approved by Human Genome Organization Gene Nomenclature Committee.

S. cerevisiae. In non-snRNP spliceosomal proteins, the majority of the human proteins have orthologs in *D. discoideum*, *A. thaliana* and *D. melanogaster* but not in *S. cerevisiae*. For the convenience of description, we list them in groups as shown in Table 3.

SR and SR-related proteins are characterized by two structural motifs, RNA recognition motif (RRM) of RNP type and RS domain containing arginine-serine rich sequences. Sixteen members of SR and SR-related proteins have been identified in human. These proteins play important roles in both constitutive splicing and alternative splicing regulation (reviewed in (Blencowe, 2000; Black, 2003; Wu et al., 2004; Sanford et al., 2005; Lin and Fu, 2007; Matlin and Moore, 2007)). Interestingly, three distinct SR protein orthologs with RRMs and an RS domain were identified in the *D. discoideum* genome. These proteins are DDB0233327, DDB0233352 and DDB0233351, corresponding to human 9G8 (SFRS7), Tra2-beta (SFRS10) and Tra2-alpha (TRA2A), respectively (Table 3A). Several classical SR proteins in mammals do not have orthologs in the *D. discoideum* genome, including SC35 and ASF/SF2 (Table S1). On the other hand, some SR protein genes in *D. discoideum* seem to have expanded in numbers. For example, two genes were identified as possible homologs of human SRp75 (SFRS4): DDB0233308 and DDB0233309. In such cases, only those with the highest level of sequence homology were included in Table 3A. It is also interesting to note that the RS domains in *D. discoideum* SR proteins appear to be more enriched in the RDR/RDRS motif rather than in the typical RS/SR sequences found in mammalian SR proteins. For example, in DDB0233308 and DDB0233327, there are long stretches of RDR/RDRS peptides, whose functional significance remains to be investigated.

All seven well-documented PRP19 complex associated proteins have orthologs in the *Dictyostelium*, human, fly, plant and yeast genomes (Table 3B). Several proteins known to act during the late stage of spliceosomal assembly and splicing

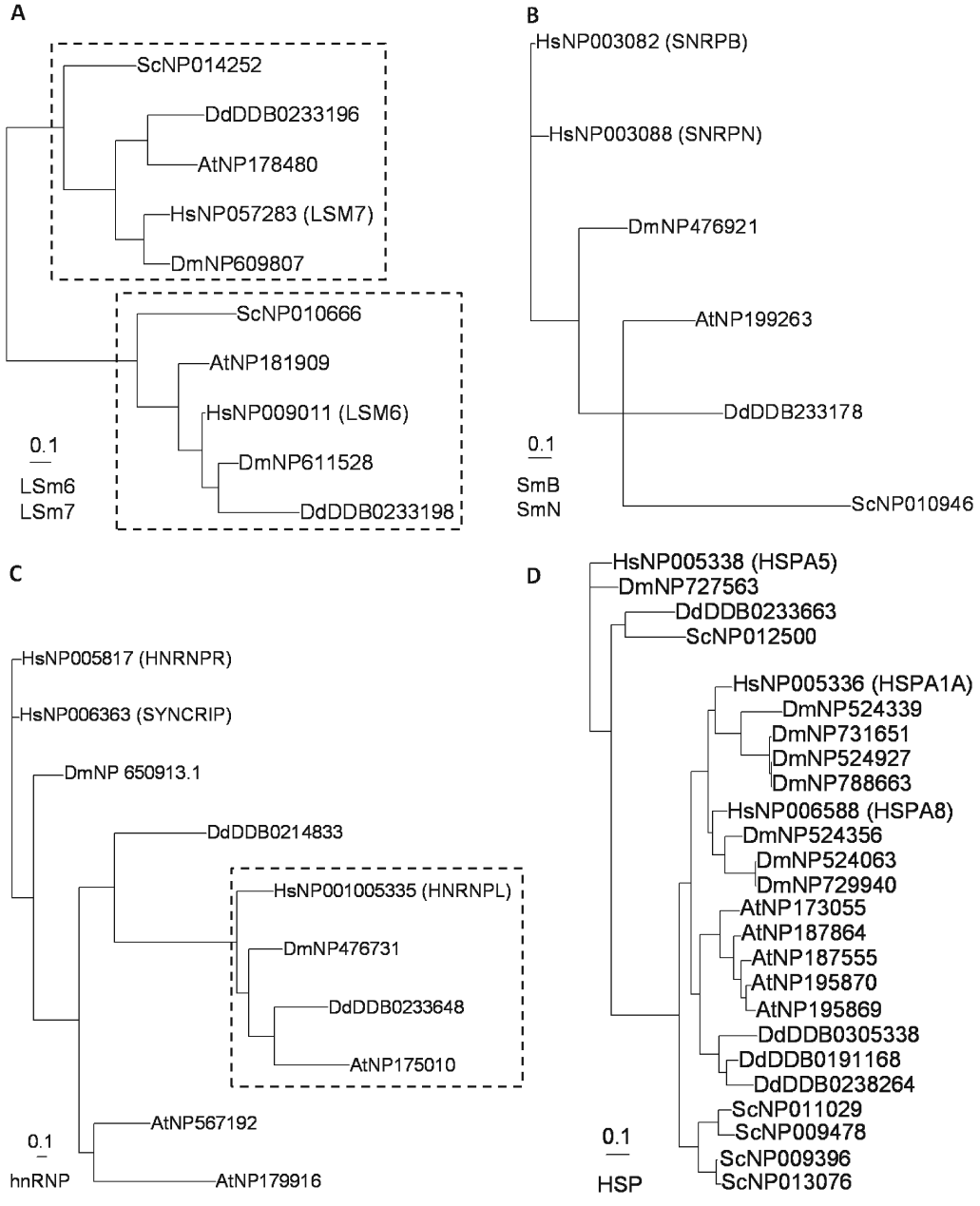

Figure 2. Phylogenetic analysis of the orthologs of spliceosomal genes in *D. discoideum*, *H. sapiens*, *D. melanogaster*, *A. thaliana* and *S. cerevisiae* genomes. Phylogenetic trees of snRNP proteins show that (A) there is the one-to-one relationship in LSm6–LSm7; and (B) only one ortholog in the fly, plant, *D. discoideum* and yeast genomes corresponds to the two human genes SNRPB (NP_003082) and SNRPN (NP_003088) in SmB–SmN. The expected evolutionary relationships are demonstrated in these conserved snRNP proteins. (C) In hnRNPs, HNRNPL (NP_001005335) has the one-to-one relationship for *D. discoideum*, plant and fly orthologs as indicated in the box, but HNRNPR and SYNCRIP have only one ortholog from *D. discoideum* and fly. No ortholog of hnRNPs was found in the yeast genome. (D) HSPA5 (NP_005338) has one-to-one ortholog from the fly, *D. discoideum* and yeast genomes. In the other two human heat shock proteins (HSP70-NP_005336, HSP71-NP_006588 and HSPA5-NP_005338), the fly genome has three orthologs to either human proteins. *D. discoideum*, plant and yeast have only a cluster of orthologs, but do not have any obvious one-to-one relationship.

were also found to be highly conserved in *Dictyostelium*, human, fly, plant and yeast, including Prp22 (DHX8), Prp43 (DHX15), Prp16 (DHX38), Slu7 (SLU7), Prp17 (CDC40) and Prp18 (PRPF18) (Table 3C).

The EJC assembly is a splicing-dependent process and serves to mark the RNA for downstream processing steps such as export, translation and nonsense-mediated decay (Tange et al., 2004; Lejeune and Maquat, 2005). The conservation of EJC proteins is high in the *D. discoideum* genome (Table 3D). Five EJC proteins corresponding to human SRRM1, BAT1, RNPS1, RBM8A and MAGOH are found in *D. discoideum*, whereas yeast has only one ortholog to human BAT1. This suggests that RNA processing could be more complex in *D. discoideum* than in yeast, although further experimental data are required for this generalization.

We identified a number of other spliceosomal proteins in *D. discoideum* that contain various motifs present in known splicing factors, including DExD, cyclophilins, WD40s, cap binding proteins, polyadenylation machinery proteins, zinc finger motif and other uncharacterized motifs (Table 3E and 3F). DExD/H containing proteins play important roles in pre-mRNA splicing (Staley and Guthrie, 1998; Cordin et al., 2006). It is interesting to note that almost all of the human spliceosomal proteins with the DExD/H motif have *D. discoideum* orthologs. Cyclophilins catalyze *cis-trans* propyl bond isomerization and facilitate protein conformational changes. All five orthologs of human splicing-related cyclophilins are present in *D. discoideum.* Searching *S. cerevisiae* with *D. discoideum* proteins identified two positive hits (NP_013633 and NP_013317; Table 3E).

hnRNP and related proteins

Forty-six heterogeneous nuclear ribonucleoproteins (hnRNPs) and other H complex associated proteins in the human genome were used to query dictyBase. Of these 46 sequences, we identified 9 non-redundant orthologs in *D. discoideum* (Table 4). The human hnRNP L (HNRNPL) hits the *D. discoideum* gene (DDB0233648) and both human hnRNP R (HNRNPR) and hnRNP Q (SYNCRIP) hit one *D. discoideum* protein (DDB0214833). None of these three hnRNPs has the yeast orthologs. This relationship was confirmed in the phylogenetic analysis (Fig. 2C). In the heat shock proteins, the human query sequences (HSPA1A and HSPA8) identified two groups of the orthologs in the fly genome, but only one cluster of the orthologs from *A. thaliana*, *D. discoideum* and *S. cerevisiae* (Fig. 2D). These clusters are not specific to either HSPA1A or HSPA8 and different from the one-to-one relationship as found in snRNPs (Fig. 2A and 2B). When these 9 non-redundant proteins were used to search the RefSeq database, all of them corresponded to the initial 16 human query sequences. The search with these putative *D. discoideum* hnRNP and related proteins also led to the identification of 15, 16 and 7 non-redundant proteins in the fly, plant and yeast genomes, respectively. In comparison with their yeast counterparts, *Dictyostelium* hnRNP protein orthologs are again more similar to those in the fly, plant and human genomes.

Alternative splicing regulators

Alternative splicing is a powerful mechanism for generating genetic diversity (More and Silver, 2008; Nilsen and Graveley, 2010).

Several groups of alternative splicing regulators have been reported in mammalian and fly genomes. These include hnRNP proteins, the SR protein super-family (SR proteins and SR-related proteins), CUGBP and ETR-like factors (CELF), DExD/H box containing proteins, RNA-binding proteins containing the heterogeneous nuclear ribonucleoprotein K-type homology (KH) or RRM domains, and other RNA binding proteins. A number of proteins are involved in both spliceosomal assembly and alternative splicing regulation.

Alternative splicing regulators of the hnRNP protein family often bind to exonic or intronic splicing regulatory sequences and influence splice site selection (reviewed in (Blencowe, 2000; Black, 2003; Wu et al., 2004; Sanford et al., 2005; Lin and Fu, 2007; Matlin and Moore, 2007)). HnRNP protein orthologs have been described in the previous section. SR proteins play important roles in both constitutive and alternative splicing (Blencowe, 2000; Cartegni et al., 2002; Wu et al., 2004; Sanford et al., 2005; Lin and Fu, 2007). Both hnRNP and SR protein orthologs have also been described in previous sections (Table 3A).

DExD/H box-containing proteins and other RNA-binding proteins also play a role in alternative splicing regulation (e.g., Wu et al., 2006; Fushimi et al., 2008; Kar et al., 2011 and references within). Two orthologs of DExD/H box containing regulators, p68 (DDX5) and p72 (DDX17), were found in *D. discoideum* (Table 3E).

The CELF family of splicing regulators interacts with CUG-containing splicing regulatory elements and control alternative splicing of a number of genes (Ladd et al., 2001). RNA transcripts containing expanded CUG/CCUG repeats can bind and sequester CUG-binding proteins and cause aberrant splicing (Ebralidze et al., 2004). Altered expression of CUG-binding proteins has been associated with myotonic dystrophy ((Kanadia et al., 2003) and reviewed in (Wang and Cooper, 2007)). Two putative CELF family members were identified in *D. discoideum* (DDB0233674 and DDB0233675), which correspond to six human CELF family members, CELF1–6 (Table 5). Three fly proteins (NP_788039, NP_609559 and NP_723739) and three plant proteins (NP_171845, NP_567249 and NP_973752) are similar to the two *D. discoideum* proteins, which are related to the above six human CELF family members (Table 5). These CELF orthologs are similar to those heat shock proteins and do not have one-to-one relationships to the human CELF proteins. No CELF proteins were found in the yeast genome.

Our sequence analyses of genomic and EST databases strongly support earlier findings (Grant and Tsang, 1990; Bain et al., 1991; Greenwood and Tsang, 1991; Escalante et al.,

2003) that *D. discoideum* has *bona fide* alternative splicing. To date, we have examined nearly all 13,527 genes individually and compared them with the available EST and cDNAs. This led to the identification of 40 genes that clearly show alternative splicing isoforms (Table 1). With only 50% of the 13,527 estimated genes in *D. discoideum* having at least some EST coverage, the actual number of alternatively spliced genes may be much higher than the 40 genes in this study. These results strongly suggest that alternative splicing could be important in the biology of this unicellular model organism. Consistent with this notion, a number of alternative splicing regulators have been identified by our sequence searches. Interestingly, all of the major families of alternative splicing regulators reported in mammals and *D. melanogaster* have been identified in *D. discoideum*. These include the SR protein super-family, CELF family, hnRNP protein family, DExD box containing proteins and other RNA binding proteins (see individual descriptions in the sections above). SR proteins are among the earliest acting proteins in spliceosome assembly. These proteins can interact with the exonic splicing regulatory elements and are related to the increased protein complexity. The CUG-binding proteins play a role in RNA processing and can regulate alternative splicing of different transcripts (Ladd et al., 2001). The expanded CUG/CCUG-containing transcripts can bind and sequester CUG-binding proteins and cause aberrant splicing (Wang and Cooper, 2007). Altered expression of CUG-binding proteins has been associated with myotonic dystrophy (Kanadia et al., 2003; Wang and Cooper, 2007). The presence of alternatively spliced genes and splicing regulators in the *D. discoideum* genome provides opportunities for studying alternative splicing in this simple model organism.

Spliceosomal snRNAs

D. discoideum snRNA genes were identified using a motif-search algorithm written in the Perl program (see METHODS section). There are five genes coding for U1 snRNAs, seven for U2 snRNAs, three for U4 snRNAs, two for U5 snRNAs, and one for U6 snRNA. Searches for U11, U12, U4atac and U6atac did not reveal convincing homologs with significant sequence similarity (data not shown), suggesting that *D. discoideum* may not have the U12 type minor class of spliceosomes. Our results of *D. discoideum* spliceosomal snRNAs are similar to the findings published by Aspegren and colleagues (Aspegren et al., 2004; Hinas et al., 2006). Taken together, it suggests that our approach can be applied in different genomes for the identification of snRNA genes.

In this study we identified 160 candidate spliceosomal proteins in the model organism *D. discoideum*. 68% of the predicted and known protein-coding genes in *D. discoideum* contain one or more introns and these genes have to undergo pre-mRNA splicing to generate functional mRNA transcripts. Therefore, pre-mRNA splicing is critical for gene expression in *D. discoideum*. In addition to all spliceosomal snRNAs (U1, U2, U4, U5 and U6), we identified 100 non-redundant sequences in the *D. discoideum* genome that are likely functional homologs of human non-snRNP spliceosomal proteins. *D. discoideum* can be used as a model system for studying the spliceosome and its components. The identification of this comprehensive set of spliceosomal proteins in *D. discoideum* should facilitate studies of pre-mRNA splicing in this model system.

The entire set of spliceosomal snRNP core proteins, the PRP19 complex proteins and late-acting splicing proteins are very highly conserved in yeast, *Dictyostelium*, plant, fly and human. Such widespread conservation suggests that these proteins play critical roles in fundamental process of pre-mRNA splicing. *D. discoideum* branches from the metazoan lineage before yeast. Our analyses show that many metazoan splicing factors that are missing in yeast are present in *D. discoideum*, indicating that these splicing-associated proteins are more ancient than previously thought. Further study will shed light on the early evolution of the metazoan splicing machinery.

Mutations in several spliceosomal protein genes, PPRC3, PRPF8 and PRPF31, cause human retinal degeneration (reviewed in (Pacione et al., 2003; Mordes et al., 2006)). It is interesting to note that all these disease-associated spliceosomal proteins are conserved in *D. discoideum*. Our comprehensive catalog of *Dictyostelium discoideum* spliceosomal proteins and related factors presented here will be useful for future experiments to elucidate splicing mechanisms and the underlying molecular pathways leading to human disease.

METHODS

Human spliceosomal proteins were collected from published studies (Hartmuth et al., 2002; Zhou et al., 2002; Wu et al., 2004; Barbosa-Morais et al., 2006; Matlin and Moore, 2007; Bessonov et al., 2008) and were used as the primary source to query dictyBase (http://www.dictybase.org/; (Chisholm et al., 2006)) using the BLASTp BLOSUM62 matrix with SEG filter (for filtering low-complexity subsequences) (Altschul et al., 1990). The *D. discoideum* protein primary features database and an E-value < 10^{-6} was used. The local alignments between the human and *D. discoideum* genes were manually reviewed to identify the structural motif regions present in the human spliceosomal proteins. Peptide sequences with only regional sequence homology but without the known motif(s) characteristic of the corresponding splicing proteins were excluded. Finally, a reciprocal BLAST search was performed using the identified *D. discoideum* hits as queries to search the RefSeq database (http://www.ncbi.nlm.nih.gov/RefSeq/). If a putative *D. discoideum* sequence matched the corresponding spliceosomal related gene in the human, fly, plant and yeast proteomes with an E-value < 10^{-5}, it was accepted as the *D. discoideum* ortholog.

When we did not identify orthologs with the above described method, dictyBase curators performed individual tblastn and blastp searches at dictyBase combined with domain analyses. As a general rule, blastp results with $\geqslant$ 25% identity over $\geqslant$ 70% overall length were considered as orthologs. This second-pass approach often identified those orthologs whose automatic gene prediction has been either incorrect or absent, which resulted in the correctly annotated genes not being present in the dictyBase primary sequence dataset. These genes were then added manually and are now publicly available. In some cases, similarity was masked by highly repetitive sequences in the *D. discoideum* gene. These are common in *D. discoideum*. In this case, blast searches and domain analyses were performed with partial deletions of repetitive strings comparing results with those obtained with the full-length protein. Phylogenetic analysis was performed to identify and confirm the ortholog proteins in different species. The protein sequences in each group were aligned using Clustal W version 2.0 (Larkin et al., 2007), to generate a character matrix in NEXUS file format. Phylogenetic analysis was then performed again on each of the protein alignments with MrBayes (Ronquist and Huelsenbeck, 2003), using Markov chain Monte Carlo to approximate the posterior probabilities of each tree. The .con file generated from MrBayes includes two consensus trees, which have been used to generate a graphical representation in the program TreeView (Page, 2002).

To identify spliceosomal snRNAs in the *D. discoideum* genome, a motif-search algorithm was applied, which was specially designed for this task and written in Perl language. In order to get evolutionarily extra-conservative short sequence segments (motifs) within snRNAs, known snRNA genes were compared among human, fly and plant (*A. thaliana)* using the ClustalW program. For every spliceosomal snRNA, sequence motifs were identified that contain nucleotide sequences identical among the three species studied. These motifs and the observed distances between them were used as input for writing our Perl program that was used to scan the *D. discoideum* genome. Additional 10%–20% sequence variations were permitted within motifs and 10%–20% length variations in distances between the motifs because some of the *D. discoideum* snRNA genes are very divergent from their animal and plant orthologs. The secondary structures of the predicted *D. discoideum* genes were examined using the M-fold program.

Alternatively spliced genes are discovered as an ongoing effort at dictyBase where each gene model is individually inspected and compared with all available EST and cDNA data.

ACKNOWLEDGEMENTS

We thank members of Wu lab for helpful suggestions and critical reading of the manuscript. This work was supported by grants to J.Y. W from NIH (EY014576 and GM070967), to A.F. from NSF Career award MCB-0643542 and to R.L.C. from NIH (GM64426 and HG02273).

ABBREVIATIONS

CELF, CUG binding protein and ETR-like factors; EJC, exon junction complex; hnRNP, heterogeneous nuclear ribonucleoprotein; RRM, RNA recognition motif; snRNA, uridine-rich small ribonucleic acid; snRNP, small ribonucleoprotein; SR protein, arginine-serine rich protein

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol 215, 403–410.

Aspegren, A., Hinas, A., Larsson, P., Larsson, A., and Söderbom, F. (2004). Novel non-coding RNAs in Dictyostelium discoideum and their expression during development. Nucleic Acids Res 32, 4646–4656.

Bain, G., Grant, C.E., and Tsang, A. (1991). Isolation and characterization of cDNA clones encoding polypeptides related to a Dictyostelium discoideum cyclic AMP binding protein. J Gen Microbiol 137, 501–508.

Baldauf, S.L., Roger, A.J., Wenk-Siefert, I., and Doolittle, W.F. (2000). A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 290, 972–977.

Barbosa-Morais, N.L., Carmo-Fonseca, M., and Aparício, S. (2006). Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion. Genome Res 16, 66–77.

Bessonov, S., Anokhina, M., Will, C.L., Urlaub, H., and Lührmann, R. (2008). Isolation of an active step I spliceosome and composition of its RNP core. Nature 452, 846–850.

Black, D.L. (2003). Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem 72, 291–336.

Blencowe, B.J. (2000). Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends Biochem Sci 25, 106–110.

Calarco, J.A., Zhen, M., and Blencowe, B.J. (2011). Networking in a global world: Establishing functional connections between neural splicing regulators and their target transcripts. RNA 17, 775–791.

Cartegni, L., Chew, S.L., and Krainer, A.R. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 3, 285–298.

Chisholm, R.L., Gaudet, P., Just, E.M., Pilcher, K.E., Fey, P., Merchant, S.N., and Kibbe, W.A. (2006). dictyBase, the model organism database for Dictyostelium discoideum. Nucleic Acids Res 34, D423–D427.

Collins, L., and Penny, D. (2005). Complex spliceosomal organization ancestral to extant eukaryotes. Mol Biol Evol 22, 1053–1066.

Cordin, O., Banroques, J., Tanner, N.K., and Linder, P. (2006). The DEAD-box protein family of RNA helicases. Gene 367, 17–37.

Crosby, M.A., Goodman, J.L., Strelets, V.B., Zhang, P., and Gelbart, W.M., and the FlyBase Consortium. (2007). FlyBase: genomes by the dozen. Nucleic Acids Res 35, D486–D491.

Ebralidze, A., Wang, Y., Petkova, V., Ebralidse, K., and Junghans, R. P. (2004). RNA leaching of transcription factors disrupts transcription in myotonic dystrophy. Science 303, 383–387.

Eichinger, L., Pachebat, J.A., Glöckner, G., Rajandream, M.A.,

Sucgang, R., Berriman, M., Song, J., Olsen, R., Szafranski, K., Xu, Q., *et al.* (2005). The genome of the social amoeba Dictyostelium discoideum. Nature 435, 43–57.

Escalante, R., Moreno, N., and Sastre, L. (2003). Dictyostelium discoideum developmentally regulated genes whose expression is dependent on MADS box transcription factor SrfA. Eukaryot Cell 2, 1327–1335.

Fushimi, K., Ray, P., Kar, A., Wang, L., Sutherland, L.C., and Wu, J.Y., (2008). Up-regulation of the proapoptotic caspase 2 splicing isoform by a candidate tumor suppressor, RBM 5. Proc Natl Acad Sci USA 105, 15708–15713.

Grant, C.E., and Tsang, A. (1990). Cloning and characterization of cDNAs encoding a novel cyclic AMP-binding protein in Dictyostelium discoideum. Gene 96, 213–218.

Greenwood, M., and Tsang, A. (1991). Sequence and expression of annexin VII of Dictyostelium discoideum. Biochim Biophys Acta 1088, 429–432.

Hartmuth, K., Urlaub, H., Vornlocher, H.-P., Will, C.L., Gentzel, M., Wilm, M., and Lührmann, R. (2002). Protein composition of human prespliceosomes isolated by a tobramycin affinity-selection method. Proc Natl Acad Sci U S A 99, 16719–16724.

Hinas, A., Larsson, P., Avesson, L., Kirsebom, L.A., Virtanen, A., and Söderbom, F. (2006). Identification of the major spliceosomal RNAs in Dictyostelium discoideum reveals developmentally regulated U2 variants and polyadenylated snRNAs. Eukaryot Cell 5, 924–934.

Hoskins, A.A., Friedman, L.J., Gallagher, S.S., Crawford, D.J., Anderson, E.G., Wombacher, R., Ramirez, N., Cornish, V.W., Gelles, J., and Moore, M.J. (2011). Ordered and dynamic assembly of single spliceosomes. Science 331, 1289–1289.

Kanadia, R.N., Johnstone, K.A., Mankodi, A., Lungu, C., Thornton, C. A., Esson, D., Timmers, A.M., Hauswirth, W.W., and Swanson, M. S. (2003). A muscleblind knockout model for myotonic dystrophy. Science 302, 1978–1980.

Kar, A., Fushimi, K., Zhou, X., Ray, P., Shi, C., Chen, X., Liu, Z., Chen, S., and Wu., J.Y., (2011). RNA helicase p68 (DDX5) regulates tau exon 10 splicing by modulating a stem-loop structure at the 5′ splice site. Mol Cell Biol 31, 1812–1821.

Ladd, A.N., Charlet, N., and Cooper, T.A. (2001). The CELF family of RNA binding proteins is implicated in cell-specific and developmentally regulated alternative splicing. Mol Cell Biol 21, 1285–1296.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., *et al.* (2007). Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947–2948.

Lejeune, F., and Maquat, L.E. (2005). Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells. Curr Opin Cell Biol 17, 309–315.

Lin, S., and Fu, X.D. (2007). SR proteins and related factors in alternative splicing. Adv Exp Med Biol 623, 107–122.

Matlin, A.J., and Moore, M.J. (2007). Spliceosome assembly and composition. Adv Exp Med Biol 623, 14–35.

Moore, M.J., and Silver, P.A., (2008). Global analysis of mRNA splicing. RNA 14, 197–203.

Mordes, D., Luo, X., Kar, A., Kuo, D., Xu, L., Fushimi, K., Yu, G., Sternberg, P. Jr, and Wu, J.Y. (2006). Pre-mRNA splicing and retinitis pigmentosa. Mol Vis 12, 1259–1271.

Nilsen, T.W., and Graveley, B.R. (2010). Expansion of the eukaryotic proteome by alternative splicing. Nature 463, 457–463.

Pacione, L.R., Szego, M.J., Ikeda, S., Nishina, P.M., and McInnes, R. R. (2003). Progress toward understanding the genetic and biochemical mechanisms of inherited photoreceptor degenerations. Annu Rev Neurosci 26, 657–700.

Page, R.D. (2002). Visualizing phylogenetic trees using TreeView. Curr Protoc Bioinformatics, Chapter 6, Unit 62.

Patel, A.A., and Steitz, J.A. (2003). Splicing double: insights from the second spliceosome. Nat Rev Mol Cell Biol 4, 960–970.

Ramani, A.K., Calarco, J.A., Pan, Q., Mavandadi, S., Wang, Y., Nelson, A.C., Lee, L.J., Morris, Q., Blencowe, B.J., Zhen, M., and Fraser, A.G., (2011). Genome-wide analysis of alternative splicing in Caenorhabditis elegans. Genome Res 21, 342–348.

Ronquist, F., and Huelsenbeck, J.P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.

Sanford, J.R., Ellis, J., and Cáceres, J.F. (2005). Multiple roles of arginine/serine-rich splicing factors in RNA processing. Biochem Soc Trans 33, 443–446.

Staley, J.P., and Guthrie, C. (1998). Mechanical devices of the spliceosome: motors, clocks, springs, and things. Cell 92, 315–326.

Tange, T.O., Nott, A., and Moore, M.J. (2004). The ever-increasing complexities of the exon junction complex. Curr Opin Cell Biol 16, 279–284.

Wang, G.S., and Cooper, T.A. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. Nat Rev Genet 8, 749–761.

Will, C.L., and Lührmann, R. (2005). Splicing of a rare class of introns by the U12-dependent spliceosome. Biol Chem 386, 713–724.

Wu, J.Y., Havlioglu, N., and Yuan, L. (2004). Alternatively spliced genes. In: Encyclopedia of Molecular Cell Biology and Molecular Medicine. Vol 1, 2nd ed. Meyers RA, ed. New York: Wiley-VCH.

Wu, J.Y., Kar, A., Kuo, D., Yu, B., and Havlioglu, N. (2006). SR_p54 (SFRS11), a regulator for tau exon 10 alternative splicing identified by an expression cloning strategy. Mol Cell Biol 26, 6739–6747.

Zhou, Z., Licklider, L.J., Gygi, S.P., and Reed, R. (2002). Comprehensive proteomic analysis of the human spliceosome. Nature 419, 182–185.

