Spliceosome Database: a tool for tracking components of the spliceosome

Ivan Cvitkovic and Melissa S. Jurica*

Department of Molecular, Cell and Developmental Biology and Center for Molecular Biology of RNA, University of California, 1156 High Street, Santa Cruz, CA 95064, USA

ABSTRACT

The spliceosome is the extremely complex macro-molecular machine responsible for pre-mRNA splicing. It assembles from five U-rich small nuclear RNAs (snRNAs) and over 200 proteins in a highly dynamic fashion. One important challenge to studying the spliceosome is simply keeping track of all these proteins, a situation further complicated by the variety of names and identifiers that exist in the literature for them. To facilitate studies of the spliceosome and its components, we created a database of spliceosome-associated proteins and snRNAs, which is available at http://spliceosomedb.ucsc.edu and can be queried through a simple browser interface. In the database, we cataloged the various names, orthologs and gene identifiers of spliceosome proteins to navigate the complex nomenclature of spliceosome proteins. We also provide links to gene and protein records for the spliceosome components in other databases. To navigate spliceosome assembly dynamics, we created tools to compare the association of spliceosome proteins with complexes that form at specific stages of spliceosome assembly based on a compendium of mass spectrometry experiments that identified proteins in purified splicing complexes. Together, the information in the database provides an easy reference for spliceosome components and will support future modeling of spliceosome structure and dynamics.

INTRODUCTION

Pre-mRNA splicing is carried out by the spliceosome, which is one of the cell's most complex and dynamic molecular machineries (1). The spliceosome assembles on each intron to be spliced from over 200 individual components (2–4). Many of these components join the spliceosome in subcomplexes, the most well known of which are the U small nuclear ribonucleoproteins (snRNPs). U snRNPs contain structured U-rich small nuclear RNAs (snRNAs) along with seven shared and several unique proteins. Assembly of U snRNPs and other proteins into the spliceosome is modeled as an ordered evolution of intermediate splicing complexes originally designated as E (early), A (pre-spliceosome), B (fully assembled) and C (catalytic). However, as new conformations of the spliceosome have been identified, additional intermediate splicing complexes (e.g. B^{act} and B^*) have been added to the assembly pathway (3,4). Likely, many more intermediate conformations of the spliceosome remain to be characterized.

The intermediate splicing complexes vary significantly in their composition, size and arrangement of components. Over the past 15 years, several research groups have used mass spectrometry peptide sequencing (MS/MS) to identify proteins that associate with different intermediate splicing complexes and subcomplexes (5–40). Each study generated lists of dozens to hundreds of proteins, and comparisons between the lists provide insight into how the spliceosome evolves between earlier and later assembly stages (2–4). For example, A complex contains U1 and U2 snRNPs and proteins involved in early recognition of the splice sites, whereas C complex contains U2, U5 and U6 snRNPs and proteins involved in promoting the second step of splicing chemistry. At different assembly stages, the spliceosomes' composition can change by well over 50 proteins, which is certainly too many for a simple mental map of the complex. The MS studies covered splicing complexes from a wide variety of species including humans, yeasts, flies and parasites. In comparing spliceosome-associated proteins between organisms, we can begin to delineate a conserved core splicing machinery, as well as potentially interesting species-specific elaborations. However, given the large number of proteins associated with the different splicing intermediate complexes and their subcomplexes, these comparisons can be challenging.

Another difficulty with comparing lists of proteins from different MS experiment is different nomenclatures. For example, while genetic studies in *Saccharomyces cerevisiae*

*To whom correspondence should be addressed. Tel: +1 831 459 4427; Fax: +1 831 459 3139; Email: jurica@biology.ucsc.edu

have been particularly important in characterizing proteins that function in splicing, only a subset of protein orthologs share the same name between yeast and humans. Furthermore, many proteins have commonly used historical names and/or multiple aliases that differ from the official gene names that have been designated by genome consortiums. One example is a subunit of the SF3a subcomplex of U2 snRNP, which was reported in different MS experiments as SAP 62, SF3a66, SF3A2 or Prp11. Another example is a Prp19 complex protein that has gone by a variety of names, including Syf1, XAB2, HCRN, Cwf3 and Ntc90. With this myriad of names, it is nearly impossible to make a straightforward comparison of the different MS analyses of splicing complexes reported in the literature.

We designed the Spliceosome Database (Spliceosome DB) with these issues in mind. It provides a simple web interface to search for spliceosome genes/proteins based on several characteristics including name(s), complex designation, identification in particular MS experiments, source organism and conserved motif/domain signatures. For each gene/protein we provide links to other databases that have amassed a great deal of additional data including links to literature, post-translational modifications, etc. Orthologous genes in several model systems are also linked. Other key features include tools for users to compare composition of different intermediate splicing complexes across several species and to directly examine the lists of proteins identified in different MS experiments. The database is a ready resource for researchers looking for information on individual spliceosome components and provides a uniquely helpful view of the dynamic assembly process of the complex.

DATABASE FEATURES

Currently, there is no simple way to query comprehensive gene and protein databases [such as Entrez (41), Ensemble (42) or UniProt (43)] to identify proteins by their association with spliceosomes exclusively. This situation makes it impossible to easily identify spliceosome proteins with specific characteristics or by their association with specific intermediate complexes. For example, a search of the Entrez gene database with "spliceosome 'C complex'" yields only three gene results and a search for "spliceosome 'B complex'" returns no hits, even though there are nearly 100 individual protein components in these splicing complexes. Likewise, it is not feasible to query within the group of spliceosome proteins by parameters such as molecular weight range, sequence motifs, domains and/or availability of structural information.

We established a database platform that mitigates these problems by cataloging spliceosome-associated components along with a variety of features. We entered components into the database based on one of the three criteria: (i) previous experimental evidence for a role in spliceosome function, (ii) homology to a known splicing factor and/or (iii) MS/MS identification of the protein product in isolated splicing complexes. Over 3600 genes/proteins from several model species, along with a variety

of key attributes, are currently recorded in the system. To allow free access from all over the world, we developed a series of web pages to query the database and return useful information. In most cases, this information can be downloaded for further off-line analysis. We envision that investigators in the pre-mRNA processing field, in particular, will utilize SpliceosomeDB for a variety of applications. In the following sections, we outline the functionalities of different database tools and their potential uses.

Searching for spliceosome components

From the 'Component Search' page of SpliceosomeDB, users can perform a quick general search of spliceosome components and their attributes or a more defined query of specific individual attributes (Figure 1A). Searchable attributes include protein and gene names and aliases, accession numbers in external databases, host organism, features of the gene's protein product such as molecular weight and conserved motifs, association with a specific snRNP or intermediate splicing complex and membership of a general protein class or family. To designate intermediate splicing complexes, we use the common E, A, B, B^{act}, C complex nomenclature associated with spliceosomes that assemble in human extracts. The general 'class/family' heading groups proteins by molecular features (e.g. SM proteins, SR proteins), association with stable spliceosome subcomplexes like the snRNPs or PRP19 complex, or other common designations (e.g. hnRNP, second step factor).

Searches result in a list of genes that match the requested parameter, and if requested, their orthologs in several model organisms. To help inspect the list, we also display basic information for the genes, including host organism, complex association, molecular weight and aliases, any of which can be sorted or filtered. Importantly, gene lists can be exported to a spreadsheet file that includes many of the individual attributes recorded in the database.

Information for individual spliceosome genes/proteins

Each gene in a search results list is linked to a page that summarizes key attributes that fall under the headings of 'Nomenclature', 'Other Resources' and 'Gene Product Info' (Figure 1B). 'Nomenclature' includes an official gene symbol, other gene symbols, full protein name and other names, all of which have been obtained from a wide variety of sources. For 'Other Resources', we provide links to several external databases including the NCBI Entrez gene database (41), UniProt protein database (43), applicable model organism-specific databases like FlyBase (44) or SGD (45), BioGrid interaction database (46), the Protein Data Bank (47,48) and SpliProt3D database of human spliceosome protein structural models recently generated by the Bujnicki Laboratory (49). 'Gene Product Info' includes molecular weight, domains, motifs, association with different intermediate splicing complexes or snRNPs and general classification.

The page also displays a list of orthologous genes from a number of model systems. Because splicing is studied in

several model organisms, it is often helpful to connect gene/protein orthologs, especially given the different naming systems employed. For most genes, we used the NCBI Homologene database (50) to identify orthologous genes. We also link orthologs to their corresponding gene page in our database and to the gene family entry in Homologene. Because Homologene does not fully cover all genes or organisms, we curated several database entries manually based on BLAST searches and extant literature.

Many pages for human genes also display a table of curated protein/protein interactions from the recent publication by the Stelzl group and recorded in SPPIR (Human Spliceosome Protein-Protein Interaction Resource) (51). The data from Hegele *et al.* (51) are derived from manual literature searches, directed yeast-two-hybrid analyses and co-precipitation experiments, and we provide links to the relevant data sources. There are other databases that provide interaction profiles for proteins, but we chose not to display these data because the interactions are not limited to spliceosome components and are not always well vetted. Often, many proteins not related to splicing appear in those lists, and although the interactions may indicate some functional linkage, they do not fully reflect what has been observed by biochemical purification and analysis of the core splicing machinery. As noted above, however, we do provide a link to BioGrid protein interaction database (46) for interested users.

Figure 1. Spliceosome 'Component Search'. (**A**) Quick Search queries all SpliceosomeDB data, whereas additional parameters can be used to limit results. The search results in a table showing matching components that can be further refined using the 'Filter current table' tool or sorted by columns. Gene/protein names are linked to a page displaying additional information about the gene/protein. Checkboxes can be used in conjunction with the 'Mass Spec Comparison' button to generate a table showing MS experiments in which the selected proteins were identified. (**B**) Portion of browser view displaying individual gene/protein information linked to additional sources of information. At the bottom of the page, MS experiments identifying the protein are listed along with the number of unique peptides by which the protein was identified.

B

Nomenclature

Gene Symbol	DHX8 (*H. sapiens*)
Other Gene Symbols	DDX8, HRH1, PRP22, PRPF22
Full Name	DEAH (Asp-Glu-Ala-His) box polypeptide 8
Other Names	PRP22, ATP-dependent RNA helicase DHX8, DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide 8 (RNA helicase), DEAH box protein 8, DEAH-box protein 8, RNA helicase HRH1

Other Resources

Entrez: Gene Info	1659
Uniprot: Protein Info	DHX8_HUMAN
BioGrid: Interaction Info	108024
PDB: Structures	2EQS, 3I4U
SpliProt3D: 3D Model	DHX8

Gene Product Info

Molecular Weight (kD)	139.3
Domains	Helicase ATP-binding, Helicase C-terminal, S1 motif
Motifs	DEAH box
Complex Association	C complex, P complex
Class / Family	Second step factors

Orthologs

Name	Organism
DHX8	G. gallus
EEED8.5	C. elegans
prp22	S. pombe
pea	D. melanogaster
PRP22	S. cerevisiae

View **Orthologs** at Homologene

Interactions
(from Hegele et al, 2012)

Filter current table:

Interacting Partner 1	Interacting Partner 2	Interaction Type	Score	PubMed ID
DHX8	CHERP	Pull-down		22365833
DHX8	CHERP	Y2H: Prey/Bait	1	22365833
DHX8	PRPF40A	Y2H: Bidirectional	1	22365833

Showing 1 to 17 of 17 entries

view interactions at SSPPIR...

Identified in the following experiments:

Export results as csv file

Filter current table:

Name	Species	# Of Peptides	PubMed ID
Prp19 tap-tag pulldown, *S. pombe*, (Ren et al 2011)	S. pombe	234	21386897
C complex on MINX-GG, *H. sapiens*, (Bessonov et al 2010)	H. sapiens	224	20980672
C complex on PM5, *H. sapiens*, (Bessonov et al 2010)	H. sapiens	211	20980672
C complex released in 1M NaCl, *H. sapiens*, (Bessonov et al 2008)	H. sapiens	201	18322460

Figure 1. Continued.

Finally, each gene page displays a list of experiments in which the encoded protein (or ortholog) was identified, usually by MS, along with the number of peptides used to identify the protein in each experiment when reported. Each MS experiment is linked to its own page, which will be described in a following section.

Displaying and comparing spliceosomal complexes

Because the spliceosome is a dynamic machine, it is important to understand its composition at the different stages of complex assembly. The point at which a component joins or leaves the spliceosome indicates a potential function in the complex. SpliceosomeDB makes it possible to quickly compare the components of different conformations of the spliceosome across several model species. From the 'Compare Complexes' page, comparisons can be made via curated component lists that we generated for different U snRNPs and spliceosome intermediate complexes. For these, we again use the complex nomenclature associated with spliceosomes that assemble in human extracts (i.e. E, A, B, B^{act}, B*, C). Users select two or more complexes and host organism, and the site will then generate a table where columns represent the selected complexes and rows display proteins present

Compare composition of spliceosome complexes

Bact complex ⇅	H. sapiens ⇅	remove
C complex ⇅	H. sapiens ⇅	remove
B complex ⇅	H. sapiens ⇅	remove
A complex ⇅	H. sapiens ⇅	remove

add another

(Compare)

Classification	A Complex H. Sapiens	B Complex H. Sapiens	Bact Complex H. Sapiens	C Complex H. Sapiens
	SNRPB	SNRPB	SNRPB	SNRPB
Recruited at B complex		SMU1		
		MFAP1		
		IK		
		WBP4		
		TFIP11		
		ZMAT2		
		PRPF38A		
Recruited at Bact complex			CWC27	CWC27
			DHX16	DHX16
			CWC22	CWC22
			ZNF830	ZNF830
			CCDC12	CCDC12
			PPIL2	
			GPKOW	GPKOW
			RNF113A	RNF113A
			PRCC	
			CWC25	
			GPATCH1	GPATCH1
Second step factors			CDC40	CDC40
				PRPF18
				SLU7
				DHX8
				SYF2
				DDX41

Figure 2. Comparing the composition of spliceosome complexes. Portion of 'Compare Complexes' browser view displaying the components of selected complexes grouped by classification. Each component is linked to its individual information page.

in those complexes, grouped by their general classification in the database (Figure 2). For comparisons between different organisms, orthologs are displayed in the same row. To designate proteins as belonging to a particular complex, we primarily drew from a recent publication by the Agafonov *et al.* (39). In that study a series of human splicing complexes were isolated under similar conditions. Associated proteins were separated by 2D electrophoresis and identified by MS/MS sequencing. The proteins were also quantified by direct staining in the gels. For each splicing complex designation, we include proteins that appear to be abundant as indicated by a high stain index. Because the association of proteins that have more dynamic interactions with the spliceosome is often not clearly stoichiometric, we also considered how

consistently the proteins have been found in spliceosome complexes at a particular stage. One such example is the SF3B subcomplex, which appears less abundant in spliceosomes captured just prior and after first step chemistry (B^{act} and C complexes) but was detected nonetheless, and we therefore include its proteins as components of both of those complexes (13,19,21,39). For these general complex association lists, we did not include proteins that appear to associate with RNA in a splicing independent manner, such as many general RNA binding proteins.

MS analyses of spliceosome complexes

An ever-growing number of studies have reported MS/MS analyses of different spliceosome complexes and subunits

(5–40). Indeed, these studies have gone far in helping to define the components of the spliceosome at different stages of assembly. In SpliceosomeDB, we recorded the results of over 135 individual MS experiments from 40 publications reporting analysis of endogenous splicing complexes isolated from cells or complexes assembled *in vitro*. The samples studied were derived from several different organisms, including human, chicken, fruit flies, yeasts and parasites. From this extensive effort by the wider splicing research community, a large portion of known splicing intermediate complexes and subunits are represented in the database. We expect that intermediates that have not yet been captured for detailed proteomic analysis will eventually succumb to analysis, and we will add those data as they are reported.

Through the 'Mass Spec Experiments' page users can find specific experiments that are identified by sample type, first author of the publication reporting the experiment or their lab head, year of publication and source organism (Figure 3A). For sample names, we use the reported assembly intermediate designation when applicable. Several samples represent less defined or mixed populations of spliceosomes, which we designated either as 'mixed-spliceosomes' or by the component target by which they were purified (i.e. 'SMD3 pulldown'). A search from this page will produce a list of matching experiments that includes an internal SpliceosmeDB id number, basic information about the sample and associated publication, along with a PubMed link to the original publication. Each experiment is linked to its own page, which displays a list of genes encoding the proteins that were identified, each of which is then linked back to its individual gene information page and, when reported, the measure of confidence/quantification associated with

Figure 3. 'Mass Spec Experiments' (**A**) MS experiments can be queried by a general 'Quick Search' or by specific attributes to return a list of matching experiments. Each experiment is linked to a page displaying the entire experimental results and to the PubMed entry of the corresponding publication. Checkboxes can be used in conjunction with the 'Mass Spec Comparison' button to generate a table comparing the results of the selected MS experiments. (**B**) Portion of browser view displaying an individual MS experiment result. Gene/protein names of identified proteins are linked to a page displaying additional information and the number of unique peptides by which proteins were identified is given.

that designation (Figure 3B). Typically, this measure is the number of unique peptides used to identify the protein, but may also refer to the total number of peptides sequenced from the protein or band intensity from gel staining analysis. For a subset of MS experiments, proteins were simply reported as present in a sample. In such cases, we use a filled box to denote the identification of the protein in the sample. To help parse the list, we also display some basic protein information including host organism, complex association, aliases and molecular weight. Again, as with most data returned by database searches, these results can be exported to a spreadsheet file.

Comparing MS experiment results

To allow users to compare results from several experiments, we created a 'Mass Spec Comparison' tool. Checkboxes are used to select a subset of genes or experiments returned from a search results returned by 'Component Search' or 'Mass Spec Experiments'. The selected items are then returned in a new page as a exportable table of genes in rows versus MS experiments in columns (Figure 4). The number of identifying peptides listed (or filled box) designates that a gene's protein product was identified in the experiment.

Documentation, discussion forum and feedback

To help users with SpliceosomeDB, the 'About' page and associated links summarize features of the database components and available tools. Also in the different search forms, definitions and examples for many of the elements and fields appear when a cursor is held over them. A 'comments/report a problem' link is available for direct feedback, and we set up a forum for moderated discussion. We welcome any comments, questions and suggestions for new features. An active dialog with users

will be important for us to identify and correct errors in the database and to keep information up to date.

DATABASE ARCHITECTURE AND WEB INTERFACE

SpliceosomeDB is backed by a MySQL relational database consisting of nine primary tables storing information for (i) individual genes/protein attributes, (ii) MS experiment details, (iii) MS results, (iv) designating protein orthologs, (v) spliceosome complexes, (vi) designating complex composition, (vii) protein classes, (viii) designating proteins to classes and (ix) taxonomy information for species represented in the database. The tables are linked as shown in Supplementary Figure S1. By leveraging relationships between tables, we have been able to cross-reference data in novel and informative ways. Data in Tables ii, iii and v–viii were entered manually. Most data in Tables i, iv and ix were downloaded from the NCBI Entrez Gene (41), UniProt (43) and NCBI Homologene (50) databases, with some attributes being manually curated.

Automatic maintenance and update is a key feature of the SpliceosomeDB. Scripts regularly populate information across database tables after the upload of new mass spectrometry experiments. Additionally, public database entries are checked for updates, with local copies stored as XML files in the event that the database schema changes, which seamlessly provides links to the previously mentioned resources. At this time, there are 3636 gene/protein entries from seven different organisms: *Homo sapiens, Gallus gallus, Drosophila melanogaster, S. cerevisiae, Schizosaccharomyces pombe, Caenorhabditis elegans, Plasmodium falicparum, Leishmania major,* and *Trypanosoma brucei.* We recorded results of 135 published MS analyses of various splicing related complexes and will add more as they are reported (5–40).

											Export results as csv file
Experiment IDs→ Proteins↓	**Classification**	**7**	**29**	**32**	**50**	**51**	**52**	**53**	**97**	**98**	**154**
SNRPB	Sm	4	9	1	26	18	15	42	31	18	257
SNRPD1	Sm	3	6	3	10	22	5	4	9	3	325
SNRPD2	Sm	3	15	3	30	68	17	27	52	34	253
SNRPD3	Sm	1	7	1	26	46	9	19	15	17	■
SNRPE	Sm	2	8	3	10	25	4		14	6	257
SNRPF	Sm	2	1		4	11	1	6	6	6	239
SNRPG	Sm	2	4		5	3	4	8	8	5	281
SNRPA	U1 snRNP								1		
SNRNP70	U1 snRNP								2	2	
SNRPA1	17S U2 snRNP	9	16	2	39	18	29	26	74	32	115
SNRPB2	17S U2 snRNP	2	9	2	13	14	4	6	30	20	133
SF3B1	17S U2 snRNP	7	4	9	22	28	48	6	75	54	29
SF3B2	17S U2 snRNP	8	8	3	16	13	10	6	31	36	35
SF3B3	17S U2 snRNP	8	12	28	26	67	42	8	105	98	44

Figure 4. Comparison of MS experiment results. Generated with the 'Mass Spec Comparison' tool, columns in the comparison table display results of individual MS experiments, typically shown as the number of unique peptides used to identify the proteins represented in the different rows. Descriptions for each MS 'Experiment ID' are displayed in a legend and linked to individual experiment results.

Web interface

A web interface for the database is located at http://spliceosomedb.ucsc.edu. The interface is built using the Django web framework with Apache serving pages. Django follows the Model, View, Template (MTV) framework of development. Data is stored in MySQL, defined by its representation (Model or schema), accessed by Python scripts called views (View) and rendered into HTML templates (Template) to be displayed through web browsers.

DISCUSSION

To aid studies of the spliceosome, which is one of the most complicated macromolecular machines in eukaryotic cells, we have established an important information source freely available to the public. Our goal was not to replicate the information already available in other databases for spliceosome components, but instead was to create a tool for navigating the complexities of spliceosome assembly dynamics and nomenclature with easy access to other information sources. To that end, SpliceosomeDB is organized from two primary perspectives: (i) information pertaining to specific spliceosome components and (ii) grouped results of mass spectrometry analysis of splicing-related samples. By allowing for cross-reference of information from these perspectives, SpliceosomeDB offers a unique and powerful tool for exploring spliceosome structure and dynamics.

Our lab has been using an in-house form of this database for several years to compare and interpret long lists of MS analyses of spliceosome complexes. From the first MS analysis of human C complex spliceosome, we identified many proteins with homologs in yeast for which genetic and biochemical studies established roles in spliceosome functions (11,18,19). However, many additional proteins with no clear ortholog in *S. cerevisiae* were also associated with human spliceosomes. Whether these proteins have a role in the spliceosome or were contaminants in our preparation was not known. By comparing results across several experimental systems, it was clear that a number of these proteins are consistently identified with splicing complexes and thereby have a higher likelihood of being *bona fide* splicing factors.

Comparisons of MS results have also been important for understanding the dynamics of spliceosome assembly and function. Differences in the composition of spliceosomes arrested at different stages of assembly likely reflect the joining and leaving of components to and from the complex, which may suggest when they function in the spliceosome. With SpliceosomeDB, other researchers are able to easily make such comparisons across a larger number of MS studies, with the ability to focus on genes of their particular interest.

One caveat of the MS data is that they are not strictly quantitative, which is to say that all proteins reported to associate with a given complex are not necessarily stoichiometric. Some clue to the relative abundance of a given protein in a complex can be derived by the number of unique peptides used to identify the protein, which a higher number indicating significant representation of the protein in the sample. However, peptide numbers also depend on protein size, with larger proteins yielding more peptides than smaller peptides. The total amount of samples analyzed and the sample complexity, also significantly affect the number of peptides sequenced for a given protein. Therefore, it is important to look at the MS data across an entire experiment to get a feel for the number of peptides that indicate likely stoichiometric presence of a protein. In that same vein, one cannot directly compare peptide numbers from different experiments, so again looking at data *en masse* is key to making judgments about the relative abundance of proteins. Fortunately, SpliceosomeDB makes it possible to display the entire data from MS experiments to provide this context.

Finally, SpliceosomeDB is useful as an organizational tool to keep track of the hundreds of spliceosome proteins and quickly find them by a number of key features. For example, we needed the list of spliceosome C complex components that are in a particular molecular weight range. Without the database, answering that question would have been very time-consuming, but with SpliceosomeDB, a straightforward query immediately returned the desired list. Furthermore, ready access to gene, protein and homolog information advances discussion with other researchers in, for example, recalling the name of a yeast homolog or association of a protein in at particular stage of spliceosome assembly. We expect that other scientists in the splicing community and beyond will also find the database useful in propagating their own studies and conversations.

Looking toward the future, we will continue to add MS data to SpliceosomeDB as they are published and plan to record additional attributes for MS experiments, such as the details of purification conditions used in isolating the samples analyzed. We welcome feedback and requests for additional features. For example, based on user input, we are currently gathering interaction data for yeast proteins including genetic interactions. SpliceosomeDB is, and will continue to be an important resource for researchers studying this complicated cellular machine.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figure 1.

REFERENCES

1. Nilsen,T.W. (2003) The spliceosome: the most complex macromolecular machine in the cell? *Bioessays*, **25**, 1147–1149.
2. Jurica,M.S. and Moore,M.J. (2003) Pre-mRNA splicing: awash in a sea of proteins. *Mol. Cell*, **12**, 5–14.
3. Wahl,M.C., Will,C.L. and Luhrmann,R. (2009) The spliceosome: design principles of a dynamic RNP machine. *Cell*, **136**, 701–718.
4. Will,C.L. and Luhrmann,R. (2011) Spliceosome structure and function. *Cold Spring Harb. Perspect. Biol.*, **3**, a003707.
5. Ajuh,P., Kuster,B., Panov,K., Zomerdijk,J.C., Mann,M. and Lamond,A.I. (2000) Functional analysis of the human CDC5L complex and identification of its components by mass spectrometry. *EMBO J.*, **19**, 6569–6581.
6. Behzadnia,N., Golas,M.M., Hartmuth,K., Sander,B., Kastner,B., Deckert,J., Dube,P., Will,C.L., Urlaub,H., Stark,H. *et al.* (2007) Composition and three-dimensional EM structure of double affinity-purified, human prespliceosomal A complexes. *EMBO J.*, **26**, 1737–1748.
7. Bessonov,S., Anokhina,M., Krasauskas,A., Golas,M.M., Sander,B., Will,C.L., Urlaub,H., Stark,H. and Luhrmann,R. (2008) Characterization of purified human Bact spliceosomal complexes reveals compositional and morphological changes during spliceosome activation and first step catalysis. *RNA*, **16**, 2384–2403.
8. Boehringer,D., Makarov,E.M., Sander,B., Makarova,O.V., Kastner,B., Luhrmann,R. and Stark,H. (2004) Three-dimensional structure of a pre-catalytic human spliceosomal complex B. *Nat. Struct. Mol. Biol.*, **11**, 463–468.
9. Carnahan,R.H., Feoktistova,A., Ren,L., Niessen,S., Yates,J.R. 3rd and Gould,K.L. (2005) Dim1p is required for efficient splicing and export of mRNA encoding lid1p, a component of the fission yeast anaphase-promoting complex. *Eukaryot Cell.*, **4**, 577–587.
10. Chen,Y.I., Moore,R.E., Ge,H.Y., Young,M.K., Lee,T.D. and Stevens,S.W. (2007) Proteomic analysis of in vivo-assembled pre-mRNA splicing complexes expands the catalog of participating factors. *Nucleic Acids Res.*, **35**, 3928–3944.
11. Coltri,P., Effenberger,K., Chalkley,R.J., Burlingame,A.L. and Jurica,M.S. (2011) Breaking up the C complex spliceosome shows stable association of proteins with the lariat intron intermediate. *PLoS One*, **6**, e19061.
12. Deckert,J., Hartmuth,K., Boehringer,D., Behzadnia,N., Will,C.L., Kastner,B., Stark,H., Urlaub,H. and Luhrmann,R. (2006) Protein composition and electron microscopy structure of affinity-purified human spliceosomal B complexes isolated under physiological conditions. *Mol. Cell. Biol.*, **26**, 5528–5543.
13. Fabrizio,P., Dannenberg,J., Dube,P., Kastner,B., Stark,H., Urlaub,H. and Luhrmann,R. (2009) The evolutionarily conserved core design of the catalytic activation step of the yeast spliceosome. *Mol. Cell*, **36**, 593–608.
14. Gottschalk,A., Neubauer,G., Banroques,J., Mann,M., Luhrmann,R. and Fabrizio,P. (1999) Identification by mass spectrometry and functional analysis of novel proteins of the yeast [U4/U6.U5] tri-snRNP. *EMBO J.*, **18**, 4535–4548.
15. Gottschalk,A., Tang,J., Puig,O., Salgado,J., Neubauer,G., Colot,H.V., Mann,M., Seraphin,B., Rosbash,M., Luhrmann,R. *et al.* (1998) A comprehensive biochemical and genetic analysis of the yeast U1 snRNP reveals five novel proteins. *RNA*, **4**, 374–393.
16. Hartmuth,K., Urlaub,H., Vornlocher,H.P., Will,C.L., Gentzel,M., Wilm,M. and Luhrmann,R. (2002) Protein composition of human prespliceosomes isolated by a tobramycin affinity-selection method. *Proc. Natl Acad. Sci. USA*, **99**, 16719–16724.
17. Herold,N., Will,C.L., Wolf,E., Kastner,B., Urlaub,H. and Luhrmann,R. (2009) Conservation of the protein composition and electron microscopy structure of Drosophila melanogaster and human spliceosomal complexes. *Mol. Cell. Biol.*, **29**, 281–301.
18. Ilagan,J., Yuh,P., Chalkley,R.J., Burlingame,A.L. and Jurica,M.S. (2009) The role of exon sequences in C complex spliceosome structure. *J. Mol. Biol.*, **394**, 363–375.
19. Jurica,M.S., Licklider,L.J., Gygi,S.R., Grigorieff,N. and Moore,M.J. (2002) Purification and characterization of native

spliceosomes suitable for three-dimensional structural analysis. *RNA*, **8**, 426–439.
20. Khanna,M., Van Bakel,H., Tang,X., Calarco,J.A., Babak,T., Guo,G., Emili,A., Greenblatt,J.F., Hughes,T.R., Krogan,N.J. *et al.* (2009) A systematic characterization of Cwc21, the yeast ortholog of the human spliceosomal protein SRm300. *RNA*, **15**, 2174–2185.
21. Lardelli,R.M., Thompson,J.X., Yates,J.R. 3rd and Stevens,S.W. (2010) Release of SF3 from the intron branchpoint activates the first step of pre-mRNA splicing. *RNA*, **16**, 516–528.
22. Luz Ambrosio,D., Lee,J.H., Panigrahi,A.K., Nguyen,T.N., Cicarelli,R.M. and Gunzl,A. (2009) Spliceosomal proteomics in Trypanosoma brucei reveal new RNA splicing factors. *Eukaryot. Cell*, **8**, 990–1000.
23. Makarov,E.M., Makarova,O.V., Urlaub,H., Gentzel,M., Will,C.L., Wilm,M. and Luhrmann,R. (2002) Small nuclear ribonucleoprotein remodeling during catalytic activation of the spliceosome. *Science*, **298**, 2205–2208.
24. Merz,C., Urlaub,H., Will,C.L. and Luhrmann,R. (2007) Protein composition of human mRNPs spliced in vitro and differential requirements for mRNP protein recruitment. *RNA*, **13**, 116–128.
25. Newo,A.N., Lutzelberger,M., Bottner,C.A., Wehland,J., Wissing,J., Jansch,L. and Kaufer,N.F. (2007) Proteomic analysis of the U1 snRNP of Schizosaccharomyces pombe reveals three essential organism-specific proteins. *Nucleic Acids Res.*, **35**, 1391–1401.
26. Ohi,M.D. and Gould,K.L. (2002) Characterization of interactions among the Cef1p-Prp19p-associated splicing complex. *RNA*, **8**, 798–815.
27. Palfi,Z., Jae,N., Preusser,C., Kaminska,K.H., Bujnicki,J.M., Lee,J.H., Gunzl,A., Kambach,C., Urlaub,H. and Bindereif,A. (2009) SMN-assisted assembly of snRNP-specific Sm cores in trypanosomes. *Genes Dev.*, **23**, 1650–1664.
28. Peng,R., Hawkins,I., Link,A.J. and Patton,J.G. (2006) The splicing factor PSF is part of a large complex that assembles in the absence of pre-mRNA and contains all five snRNPs. *RNA Biol.*, **3**, 69–76.
29. Rappsilber,J., Ryder,U., Lamond,A.I. and Mann,M. (2002) Large-scale proteomic analysis of the human spliceosome. *Genome Res.*, **12**, 1231–1245.
30. Sharma,S., Falick,A.M. and Black,D.L. (2005) Polypyrimidine tract binding protein blocks the 5′ splice site-dependent assembly of U2AF and the prespliceosomal E complex. *Mol. Cell*, **19**, 485–496.
31. Sharma,S., Kohlstaedt,L.A., Damianov,A., Rio,D.C. and Black,D.L. (2008) Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome. *Nat. Struct. Mol. Biol.*, **15**, 183–191.
32. Stevens,S.W., Barta,I., Ge,H.Y., Moore,R.E., Young,M.K., Lee,T.D. and Abelson,J. (2001) Biochemical and genetic analyses of the U5, U6, and U4/U6 x U5 small nuclear ribonucleoproteins from Saccharomyces cerevisiae. *RNA*, **7**, 1543–1553.
33. Stevens,S.W., Ryan,D.E., Ge,H.Y., Moore,R.E., Young,M.K., Lee,T.D. and Abelson,J. (2002) Composition and functional characterization of the yeast spliceosomal penta-snRNP. *Mol. Cell*, **9**, 31–44.
34. Tkacz,I.D., Gupta,S.K., Volkov,V., Romano,M., Haham,T., Tulinski,P., Lebenthal,I. and Michaeli,S. (2010) Analysis of spliceosomal proteins in Trypanosomatids reveals novel functions in mRNA processing. *J. Biol. Chem.*, **285**, 27982–27999.
35. Wang,Q., Hobbs,K., Lynn,B. and Rymond,B.C. (2003) The Clf1p splicing factor promotes spliceosome assembly through N-terminal tetratricopeptide repeat contacts. *J. Biol. Chem.*, **278**, 7875–7883.
36. Will,C.L., Schneider,C., Hossbach,M., Urlaub,H., Rauhut,R., Elbashir,S., Tuschl,T. and Luhrmann,R. (2004) The human 18S U11/U12 snRNP contains a set of novel proteins not found in the U2-dependent spliceosome. *RNA*, **10**, 929–941.
37. Zhang,C., Dowd,D.R., Staal,A., Gu,C., Lian,J.B., van Wijnen,A.J., Stein,G.S. and MacDonald,P.N. (2003) Nuclear coactivator-62 kDa/Ski-interacting protein is a nuclear matrix-associated coactivator that may couple vitamin D receptor-mediated transcription and RNA splicing. *J. Biol. Chem.*, **278**, 35325–35336.

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

39. Agafonov,D.E., Deckert,J., Wolf,E., Odenwalder,P., Bessonov,S., Will,C.L., Urlaub,H. and Luhrmann,R. (2011) Semiquantitative proteomic analysis of the human spliceosome via a novel two-dimensional gel electrophoresis method. *Mol. Cell. Biol.*, **31**, 2667–2682.

40. Makarov,E.M., Owen,N., Bottrill,A. and Makarova,O.V. (2011) Functional mammalian spliceosomal complex E contains SMN complex proteins in addition to U1 and U2 snRNPs. *Nucleic Acids Res.*, **40**, 2639–2652.

41. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **39**, D52–D57.

42. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al.* (2012) Ensembl 2012. *Nucleic Acids Res.*, **40**, D84–D90.

43. Magrane,M. and Consortium,U. (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database*, **2011**, bar009.

44. McQuilton,P., St Pierre,S.E. and Thurmond,J. (2012) FlyBase 101–the basics of navigating FlyBase. *Nucleic Acids Res.*, **40**, D706–D714.

45. Cherry,J.M., Hong,E.L., Amundsen,C., Balakrishnan,R., Binkley,G., Chan,E.T., Christie,K.R., Costanzo,M.C., Dwight,S.S., Engel,S.R. *et al.* (2012) Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.*, **40**, D700–D705.

46. Stark,C., Breitkreutz,B.J., Chatr-Aryamontri,A., Boucher,L., Oughtred,R., Livstone,M.S., Nixon,J., Van Auken,K., Wang,X., Shi,X. *et al.* (2011) The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res.*, **39**, D698–D704.

47. Bernstein,F.C., Koetzle,T.F., Williams,G.J., Meyer,E.F. Jr, Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The Protein Data Bank. A computer-based archival file for macromolecular structures. *Eur. J. Biochem.*, **80**, 319–324.

48. Rose,P.W., Beran,B., Bi,C., Bluhm,W.F., Dimitropoulos,D., Goodsell,D.S., Prlic,A., Quesada,M., Quinn,G.B., Westbrook,J.D. *et al.* (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.*, **39**, D392–D401.

49. Korneta,I., Magnus,M. and Bujnicki,J.M. (2012) Structural bioinformatics of the human spliceosomal proteome. *Nucleic Acids Res.*, **40**, 7046–7065.

50. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Federhen,S. *et al.* (2012) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **40**, D13–D25.

51. Hegele,A., Kamburov,A., Grossmann,A., Sourlis,C., Wowro,S., Weimann,M., Will,C.L., Pena,V., Luhrmann,R. and Stelzl,U. (2012) Dynamic protein-protein interaction wiring of the human spliceosome. *Mol. Cell*, **45**, 567–580.