Published online 15 November 2006 *Nucleic Acids Research, 2007, Vol. 35, Database issue* ***D93–D98***
doi:10.1093/nar/gkl884

The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species

Namshin Kim, Alexander V. Alekseyenko[1], Meenakshi Roy and Christopher Lee*

Department of Chemistry and Biochemistry, Center for Computational Biology, Institute for Genomics and Proteomics, Molecular Biology Institute, University of California Los Angeles, Los Angeles, CA 90095-1570, USA and [1]Department of Biomathematics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA 90095, USA

Received September 15, 2006; Revised October 10, 2006; Accepted October 11, 2006

ABSTRACT

We have greatly expanded the Alternative Splicing Annotation Project (ASAP) database: (i) its human alternative splicing data are expanded ~3-fold over the previous ASAP database, to nearly 90 000 distinct alternative splicing events; (ii) it now provides genome-wide alternative splicing analyses for 15 vertebrate, insect and other animal species; (iii) it provides comprehensive comparative genomics information for comparing alternative splicing and splice site conservation across 17 aligned genomes, based on UCSC multigenome alignments; (iv) it provides an ~2- to 3-fold expansion in detection of tissue-specific alternative splicing events, and of cancer versus normal specific alternative splicing events. We have also constructed a novel database linking orthologous exons and orthologous introns between genomes, based on multigenome alignment of 17 animal species. It can be a valuable resource for studies of gene structure evolution. ASAP II provides a new web interface enabling more detailed exploration of the data, and integrating comparative genomics information with alternative splicing data. We provide a set of tools for advanced data-mining of ASAP II with Pygr (the Python Graph Database Framework for Bioinformatics) including powerful features such as graph query, multigenome alignment query, etc. ASAP II is available at http://www.bioinformatics.ucla.edu/ASAP2.

INTRODUCTION

Alternative splicing plays an important role in protein diversity and gene regulation (1–3). Recent studies on alternative splicing estimate that 40–70% of human genes are alternatively spliced (4–6). Moreover, many splice variants alter the function of the protein product, and are involved in human diseases (7). Thus, alternative splicing is an important medical target for development of novel diagnostics and therapeutic drugs (8).

Genome-wide analyses of alternative splicing are mainly based on publicly available sequence databases such as GenBank (9) and Swiss-Prot/TrEMBL. HOLLYWOOD (10) and ASD (11) give comprehensive analyses of alternative splicing for human and mouse. Notably, those two databases provide with comparative studies between human and mouse. Lee *et al*. (12) constructed DEDB for genome-wide analysis of alternative splicing for *Drosophila melanogaster*. As well as alternative splicing analysis, ECgene (13,14) gives comprehensive analysis results for functional annotation of proteins and expression analysis. Furthermore, it has been recently expanded to nine species.

The Alternative Splicing Annotation Project (ASAP) database (15) is a widely used resource providing a genome-wide analysis of human alternative splicing and tissue-specific splicing (4,16–20) based on expressed sequence tag (EST), messenger RNA (mRNA) and genome sequences. It has served as the basis for a wide variety of studies (21–28).

Here we describe a major expansion of the ASAP database, designed to make it a good resource for analyzing and comparing alternative splicing between a wide range of animal genomes. Whereas the original release of ASAP focused entirely on human data, we have now included genome-wide analyses of alternative splicing for 15 animal species from human to nematodes. Furthermore, we have added a new dimension of comparative genomics tools, for comparing alternative splicing patterns, conservation of splice sites, exons and introns, across 17 animal genomes.

MATERIALS AND METHODS

We downloaded UniGene (29), GenBank (9) and Entrez Genes (30) from NCBI ftp site (UniGene; ftp://ftp.ncbi.nih.gov/repository/UniGene/, GenBank; ftp://ftp.ncbi.nih.gov/genbank/, Entrez Genes;ftp://ftp.ncbi.nih.gov/gene/) in January 2006. Genome assembly sequences, RefSeq (31)/mRNA alignments and RepeatMasker tracks were downloaded from UCSC genome browser except for yellow fever mosquito genome from Ensembl genome browser

*To whom correspondence should be addressed. Tel: +1 310 825 7374; Fax: +1 310 206 7286; Email: leec@mbi.ucla.edu

(32). Multigenome alignments for human (hg17), mouse (mm7), chicken (galGal2), fruit fly (dm2), zebrafish (danRer3) and western clawed frog (xenTro1) were downloaded from UCSC genome browser.

In order to update lists of tissue and cancer versus normal specific genes for human, we downloaded EST library information from UniLib (ftp:/ftp.ncbi.nih.gov/repository/UniLib/). A total of 2895 new human EST libraries were classified and added into existing 47 tissue categories and normal/tumor types. In total, 8828 human EST libraries were classified into 47 tissues and normal/tumor. We used same method used by Xu and Lee (19) for LOD value calculation for tissue and normal versus cancer specificity.

Orthologous exons, introns and splice site sequences were extracted using Pygr, which gives us less than a millisecond access to any location of any genome in multigenome alignments. Moreover, Pygr can be easily integrated with ASAP II database and more detailed information will be available at ASAP II website.

We defined as orthologous exons and introns if at least one of the splice sites of exons (those of flanking exons for introns) from two species is exactly aligned in multigenome alignments. This strategy can increase the possibilities of finding orthologous exons, because the exons can be within well-conserved blocks of multigenome alignments. Conventional protein similarity-based method can give only orthologous genes only if protein sequences are available. Moreover, multigenome alignment-based method enables us to interpret how alternatively spliced exons and introns are evolved across distant species.

RESULTS AND DISCUSSION

Alternative splicing analyses

Compared with the previous release of ASAP (15), ASAP II provides an ~3-fold expansion in human alternative splicing events, to a total of 89 078 distinct alternative splicing relationships in human, detected within 11 717 genes (UniGene clusters). Out of the total set of multi-exon genes (22 220), 53% were detected to contain alternative splicing (Table 1). Focusing on genes with at least one mRNA sequence (for which our gene model is therefore likely to be full-length, and which generally have higher EST coverage), 75% (10 202 out of 13 690) were detected to contain alternative splicing. The continuing rapid growth in alternative splicing detection as a function of increased EST and mRNA counts suggests that the field is still far from saturation, and that far more experimental data will be required to obtain a complete catalog of human alternative splicing.

Another major change is the addition of alternative splicing analyses for 14 new animal genomes (Table 1), ranging from mammals, birds and fish, to insects, *C.elegans* and *Ciona*. This provides a very large dataset of non-human alternative splicing events (a total 67 095 alternative splicing relationships, over three-quarters the size of the human alternative splicing dataset). However, due to the limited EST coverage for many animal genomes (e.g. *Fugu*, honeybee), these data cannot be considered comprehensive. Numbers of mapped UniGene clusters can be lower than expected for *Ciona, Fugu* and yellow fever mosquito due to the incomplete genome assemblies. For mouse, 8711 (53%) out of 16 404 multi-exon genes were detected to contain alternative splicing and 60% (8203 out of 13 626) for genes with at least one mRNA. Twenty five percent of Rat, 22% of western clawed frog, 22% of chicken, 26% of cow and 19% of fruit fly multi-exon genes were detected to contain alternative splicing. Proportions of the alternatively spliced multi-exon genes for *C.elegans* (6%) and African malaria mosquito (8%) were lower than mammals. Alternative splicing analyses of 15 most sequenced species can expand our research area from human to nematodes as

Table 1. Statistics for ASAP II database

Organism	Genome assembly[a]	UniGene clusters		Detected splices/Clusters		Isoforms	Alternative splicing		Alternatively spliced[b] (%)
		Total	Mapped	Splices	Clusters		Relationships	Clusters	
Human	hg17	66 488	47 477	193 023	22 220	260 198	89 078	11 717	53
Mouse	mm7	43 104	32 522	141 284	16 404	135 465	33 057	8711	53
Rat	rn3	41 687	34 003	82 941	14 195	53 212	7210	3378	24
Western clawed frog	xenTro1	33 132	24 617	65 633	10 880	34 293	4836	2349	22
Chicken	galGal2	30 470	19 708	51 471	9671	26 557	4244	2154	22
Cow	bosTau2	39 432	28 709	60 813	11 448	32 401	6692	3008	26
Dog	canFam2	22 930	16 645	29 290	6834	11 424	1633	951	14
C.elegans	ce2	20 621	15 546	54 395	12 580	23 393	1309	763	6
Ciona	ci2	15 587	1373	5611	972	2161	150	98	10
Zebrafish	danRer3	32 400	22 297	67 598	12 136	27 547	2611	1577	13
Fruit fly	dm2	16 635	14 568	37 469	9683	26 854	4850	1841	19
Fugu	fr1	2355	1980	3014	798	866	33	24	3
Yellow fever mosquito	AaegL 1	15 182	10 624	3594	1787	2529	120	87	5
Honeybee	apiMel2	5900	5027	6270	2548	2990	90	57	2
African malaria mosquito	anoGam1	15 609	14 173	17 278	8013	15 115	1070	605	8

[a] Genome assembly sequences were downloaded from UCSC genome browser except for Yellow fever mosquito, which was downloaded from Enesmbl genome browser.

[b] Alternative spliced genes (%) = No. of alternatively spliced clusters/No. of spliced clusters.

Fifty-three percent of human and mouse multi-exon genes are detected to contain alternative splicing. Focusing on genes with at least mRNA, 75 and 60% of human and mouse multi-exons genes were detected to contain alternative splicing (ASAP II website for details). Due to limited mRNA and EST coverage (*Fugu* and honeybee) and incomplete genome assembly (*Fugu*, *Ciona* and yellow fever mosquito), number of mapped clusters (*Ciona*, 9%; 1373 out of 15 587) or alternatively spliced clusers (24 for *Fugu*, 98 for *Ciona*, 57 for honeybee, 87 for and yellow fever mosquito) can be significantly lower than expected, these data cannot be considered comprehensive: 19–26% of fruit fly, western clawed frog, chicken, rat and cow multi-exon genes were detected to contain alternative splicing.

well as comparative and evolutionary studies between distantly related species.

As an illustration of ASAP II's value for biological discovery, we performed analyses of tissue-specificity and cancer versus normal specificity of human alternative splice forms. ASAP II yielded ∼2- to 3-fold larger identification of tissue-specific splice forms than the previous ASAP release (19,20). We added 2895 new EST libraries to our tissue classification database (Materials and Methods): each library source was classified as one of 47 tissue types, and also as tumor versus normal in origin. We found 1709 high-confidence (LOD $\geqslant 3$) tissue-specific alternative splicing relationships from 960 genes, and 273 high-confidence (LOD $\geqslant 3$) cancer-specific relationships from 198 genes. The largest categories of tissue-specific splice forms were identified from brain/nerve, testis, skin, muscle and lymph. Users can download all EST library classification and log-odds (LOD) calculation results from ASAP II download page and mine their own experimental candidates.

Comparative genomics analyses

To help researchers easily compare alternative splicing data between species, we performed a comprehensive comparative genomics analysis across 17 genomes (Table 2), identifying orthologous exons, introns and alternative splice events between these genomes. As a separate analysis that is valid even when the target genome has little or no alternative splicing data, we also analyzed the conservation of alternative exons and splice sites across 17 genomes. To do this, we used the well-established and characterized multigenome alignments (33) constructed for the UCSC genome browser (34). Orthologous exons and introns were defined by sharing at least one splice site in multigenome alignments (Materials and Methods). Out of 129 981, 85 673 (66%) human internal exons have at least one orthologous exon, which are identified by hg17 referenced 17 species multigenome alignments. Total numbers of orthologous exons found by five different multigenome alignments are summarized in Table 2. This method can give more comprehensive database for orthologous genes than conventional protein similarity-based method. Furthermore, we constructed multigenome splice site database from UCSC multigenome alignments (Figure 1D). These data give users both the ability to compare observed splicing patterns between experimental data for different species, but also to study the evolution of alternative exons and splice sites (by looking at their conservation) even in genomes for which no splicing data are available.

Table 2. Statistics for orthologous exons and introns

Multiple alignments[a]	Exons with orthologous exons	Total internal exons	Introns with orthologous introns	Total canonical introns
hg17 referenced 17 species Multigenome alignments	85 673	129 981	100 447	193 024
mm7 referenced 17 species Multigenome alignments	81 296	105 260	97 371	141 285
galGal2 referenced 7 species Multigenome alignments	20 471	36 865	24 973	51 472
danRer3 referenced 5 species Multigenome alignments	18 977	50 792	22 367	67 599
xenTro1 referenced 5 species Multigenome alignments	23 428	49 679	26 893	65 634

[a] Only orthologous exons and introns that have two exact matches of both canonical splice sites (U1/U2 and U11/U12). List of species used in multigenome alignments is available at UCSC genome browser (34).

Sixty-six percent (85 673 out of 1 29 981) for human and 77% (81 296 out of 105 260) for mouse internal exons have at least one orthologous exons; 52% (1 00 447 out of 1 93 024) for human and 69% (97 371 out of 1 41 285) for mouse canonical introns have at least one orthologous introns. Most of orthologous exons and introns were from human and mouse orthologs due to larger number of mRNA and EST sequences than other genomes: 56, 37 and 47% of chicken (galGal2), zebrafish (danRer3), and western clawed frog (xenTro1) internal exons have at least one orthologous exons and 49, 33 and 41% for orthologous introns. Because a set of genome assemblies used for multigenome alignments is different from ASAP II calculation for chicken, zebrafish and western clawed frog (Table 1 for details), numbers of orthologous exons and intron can be decreased.

Database mining and tools

Users can mine ASAP II in several ways:

(i) by using the web interface (below);
(ii) by downloading it as MySQL tables and performing SQL queries;
(iii) by using Python tools that work directly with the ASAP II schema, for graph query of alternative splicing graphs and comparative genomics query of multigenome alignments.

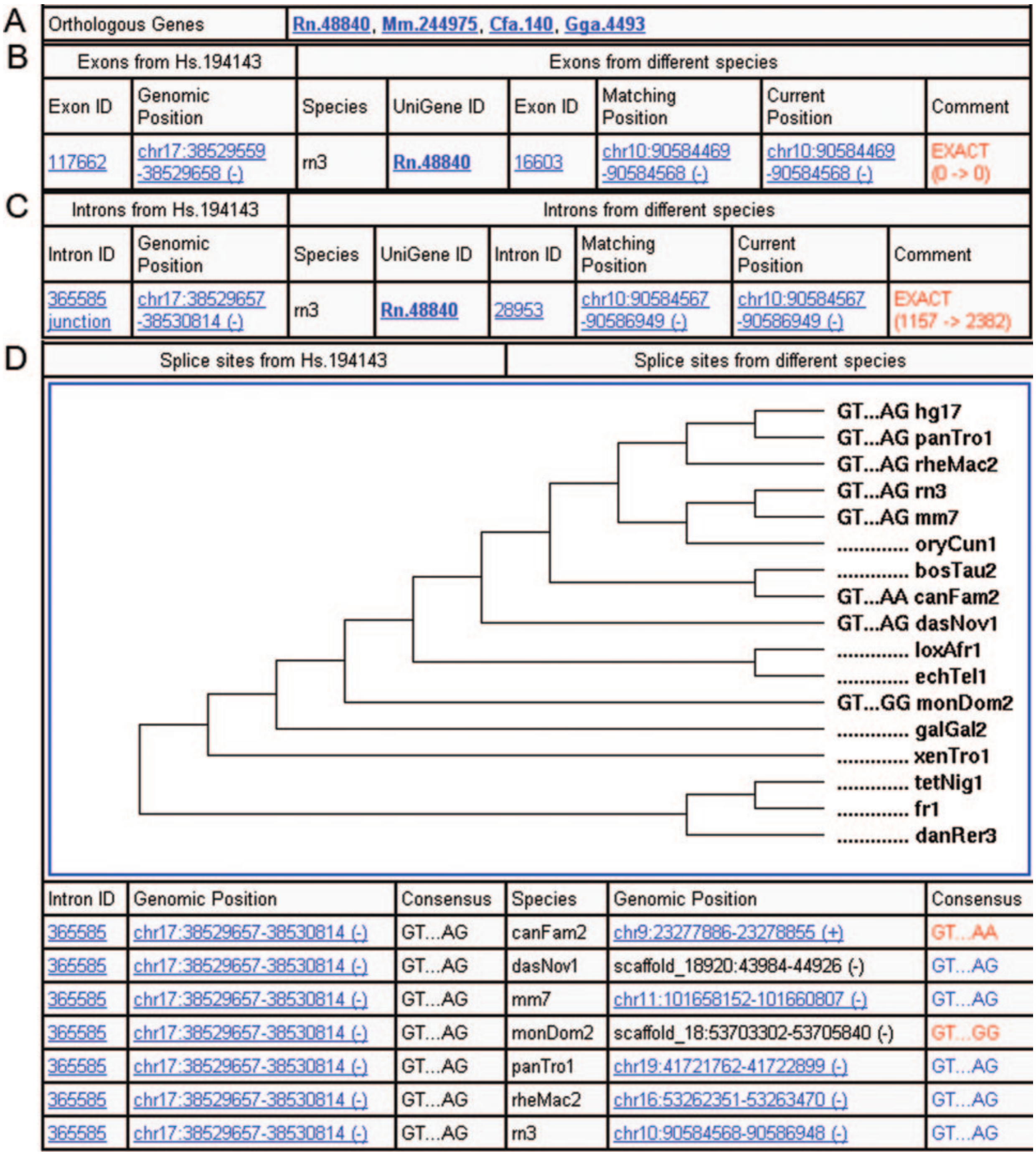
Although there's no space to discuss the latter tools (Pygr, the Python Graph Database Framework for Bioinformatics) here, extensive documentation is available on the web (http://www.bioinformatics.ucla.edu/pygr), including many tutorial examples about mining ASAP II.

Web interface

ASAP II can be searched by several different criteria such as gene symbol, gene name and ID [UniGene (29), GenBank (9), etc.]. The web interface provides seven different kinds of views:

(i) user query, UniGene annotation, orthologous genes and genome browsers;
(ii) genome alignment;
(iii) exons & orthologous exons;
(iv) introns & orthologous introns;
(v) alternative splicing;
(vi) isoform and protein sequences;
(vii) tissue & cancer versus normal specificity.

ASAP II shows genome alignments of isoforms, exons and introns in UCSC-like genome browser. Users can easily navigate among all the views by clicking links of interest. Alternative and constitutive exons are highlighted in red and blue, respectively. All alternative splicing relationships with supporting evidence information, types of alternative splicing patterns, and inclusion rate for skipped exons are listed in separate tables. Users can also search human data for tissue- and cancer-specific splice forms at the bottom of the gene summary page. We report *P*-values for tissue-specificity as



Orthologous Genes	Rn.48840, Mm.244975, Cfa.140, Gga.4493

Exons from Hs.194143		Exons from different species					
Exon ID	Genomic Position	Species	UniGene ID	Exon ID	Matching Position	Current Position	Comment
117662	chr17:38529559 -38529658 (-)	rn3	Rn.48840	16603	chr10:90584469 -90584568 (-)	chr10:90584469 -90584568 (-)	EXACT (0 -> 0)

Introns from Hs.194143		Introns from different species					
Intron ID	Genomic Position	Species	UniGene ID	Intron ID	Matching Position	Current Position	Comment
365585 junction	chr17:38529657 -38530814 (-)	rn3	Rn.48840	28953	chr10:90584567 -90586949 (-)	chr10:90584567 -90586949 (-)	EXACT (1157 -> 2382)

Splice sites from Hs.194143	Splice sites from different species

Intron ID	Genomic Position	Consensus	Species	Genomic Position	Consensus
365585	chr17:38529657-38530814 (-)	GT...AG	canFam2	chr9:23277886-23278855 (+)	GT...AA
365585	chr17:38529657-38530814 (-)	GT...AG	dasNov1	scaffold_18920:43984-44926 (-)	GT...AG
365585	chr17:38529657-38530814 (-)	GT...AG	mm7	chr11:101658152-101660807 (-)	GT...AG
365585	chr17:38529657-38530814 (-)	GT...AG	monDom2	scaffold_18:53703302-53705840 (-)	GT...GG
365585	chr17:38529657-38530814 (-)	GT...AG	panTro1	chr19:41721762-41722899 (-)	GT...AG
365585	chr17:38529657-38530814 (-)	GT...AG	rheMac2	chr16:53262351-53263470 (-)	GT...AG
365585	chr17:38529657-38530814 (-)	GT...AG	rn3	chr10:90584568-90586948 (-)	GT...AG

Figure 1. Popup page for orthologous exons, introns and splice sites. (**A**) List of orthologous genes are described in UniGene summary section. (**B**) Orthologous Exons. 'EXACT' means both splice sites are exactly aligned in multigenome alignments. Change in protein modularity (remainder divided by three) is denoted as '(0 → 0)'. (**C**) Orthologous Introns. '(1157 → 2382)' means human intron (ID 365585) size is increased in rat intron (ID 28953). (**D**) Multiple alignments of splice site sequences and its phylogenetic tree generated by UCSC Phylogenetic Tree Gif Maker (34) on the fly. Splice site consensus of this intron (ID 365585) is well-conserved within close genomes but not in distant genomes; dog (canFam2) and opossum (monDom2).

LOD scores, and highlight the results for LOD ⩾ 3 and at least three EST sequences (19,20). A short introduction to the web interface and a comprehensive user guide are available at the ASAP II website, http://www.bioinformatics.ucla.edu/ASAP2.

Comparative genomics is a major focus of the ASAP II web interface, displaying results from its new orthologous exons and introns database. For example, it displays the multiple alignments of splice site sequences as a phylogenetic tree (Figure 1D), enabling users to infer the evolutionary history of introns at a glance. In Figure 1D, one can easily that this pair of splice sites appears to have evolved in an early mammalian ancestor, but not before. Many applications are possible. For example, researchers could identify 'recently

Table 3. Comparison of alternative splicing analyses with other databases

Database	Species	Genes			Internal exons				
		Alternatively spliced		Spliced	Constitutive		Alternative		Total
ASD[a]	Human	9929	61%	16 293	—	—	—	—	—
	Mouse	8211	50%	16 391	—	—	—	—	—
ECgene[a]	Human	21 266	43%	49 546	—	—	—	—	—
	Mouse	17 706	40%	43 932	—	—	—	—	—
	Rat	8699	32%	27 406	—	—	—	—	—
DEDB[a]	Fruit fly	2646	20%	13 222	—	—	—	—	—
HOLLYWOOD[a]	Human	—	—	—	114 839	75%	37 366	25%	151 199
	Mouse	—	—	—	79 217	87%	11 673	13%	90 885
ASAP II	Human	11 717	53%	22 220	83 193	64%	46 788	36%	129 981
	Mouse	8711	53%	16 404	82 839	79%	22 421	21%	105 260
	Rat	3378	24%	14 195	—	—	—	—	—
	Fruit fly	1841	19%	9683	—	—	—	—	—

[a]All numbers are taken from ASD (11), ECgene (14), DEDB (12) and HOLLYWOOD (10).

evolved splice sites' by selecting introns whose canonical splice site sequences (GT/AG) are only conserved within closely related species, but not in distant species. ASAP II includes links to comparative genomics information from all views. All orthologous genes identified by multigenome alignments are listed in its annotation summary (Figure 1A). If the user clicks 'Show Orthologous Exons/Introns' on any page, detailed information will be shown in new window (Figure 1B and C).

Comparison with other alternative splicing databases

Alternative splicing analysis results can be significantly different between different databases because each database uses different sequence databases, genome assembly, methods for sequence alignments, alignment filtering and stringency, etc. Total numbers of alternatively spliced genes and exons for other databases are summarized in Table 3. ASAP II has more alternatively spliced genes than ASD for human (11 717 versus 9929) and mouse (8711 versus 8211). But, DEDB has more spliced genes than ASAP II (13 222 versus 9683). ECgene has twice as many spliced genes as the other databases suggesting the use of different stringency criteria for alignment filtering. HOLLYWOOD has more human internal exons than ASAP II (151 199 versus 129 981), but percentage of alternative exons is significantly lower for human (25% versus 36%) and mouse (13% versus 21%). Presumably, sequence database for HOLLYWOOD (January 2004) is older than ASAP II (January 2006).

Update and future directions

ASAP II gives alternative splicing analysis of UniGene data released in January 2006 (Version JAN06). In order to provide with up-to-date alternative splicing analysis, ASAP II database will be updated within 2 years if total number of available sequences are significantly increased. Availability of genome assembly is essential for supporting new species; we will add new species if the genome assembly is publicly available as well as the orthologous Exon/Intron database.

We will also develop novel analysis methods for alternative splicing such as evolutionary history of exons and introns and make available in ASAP II. We hope that ASAP II can become a useful resource for comparative genomics studies in the post-genome era.

ACKNOWLEDGEMENTS

The authors wish to thank Calvin Pan, Qi Wang and Dr Yi Xing for valuable comments on this work. This work has been supported by NIH grant U54 RR021813, Department of Energy grant DE-FC02-02ER63421, and by a Dreyfus Foundation Teacher-Scholar Award to C.J.L. Funding to pay the Open Access publication charges for this article was provided by NIH grant U54 RR021813.

Conflict of interest statement. None declared.

REFERENCES

1. Black,D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.*, **72**, 291–336.
2. Graveley,B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.*, **17**, 100–107.
3. Maniatis,T. and Tasic,B. (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature*, **418**, 236–243.
4. Modrek,B., Resch,A., Grasso,C. and Lee,C. (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.*, **29**, 2850–2859.
5. Kan,Z., States,D. and Gish,W. (2002) Selecting for functional alternative splices in ESTs. *Genome Res.*, **12**, 1837–1845.
6. Johnson,J.M., Castle,J., Garrett-Engele,P., Kan,Z., Loerch,P.M., Armour,C.D., Santos,R., Schadt,E.E., Stoughton,R. and Shoemaker,D.D. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **302**, 2141–2144.
7. Caceres,J.F. and Kornblihtt,A.R. (2002) Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet.*, **18**, 186–193.
8. Mangasarian,A. (2005) Alternative RNA splicing and drug target identification. *IDrugs*, **8**, 725–729.
9. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2006) GenBank. *Nucleic Acids Res.*, **34**, D16–D20.
10. Holste,D., Huo,G., Tung,V. and Burge,C.B. (2006) HOLLYWOOD: a comparative relational database of alternative splicing. *Nucleic Acids Res.*, **34**, D56–D62.
11. Stamm,S., Riethoven,J.J., Le Texier,V., Gopalakrishnan,C., Kumanduri,V., Tang,Y., Barbosa-Morais,N.L. and Thanaraj,T.A. (2006) ASD: a bioinformatics resource on alternative splicing. *Nucleic Acids Res.*, **34**, D46–D55.
12. Lee,B.T., Tan,T.W. and Ranganathan,S. (2004) DEDB: a database of Drosophila melanogaster exons in splicing graph form. *BMC Bioinformatics*, **5**, 189.

13. Kim,P., Kim,N., Lee,Y., Kim,B., Shin,Y. and Lee,S. (2005) ECgene: genome annotation for alternative splicing. *Nucleic Acids Res*., **33**, D75–D79.
14. Kim,N., Shin,S. and Lee,S. (2005) ECgene: genome-based EST clustering and gene modeling for alternative splicing. *Genome Res*., **15**, 566–576.
15. Lee,C., Atanelov,L., Modrek,B. and Xing,Y. (2003) ASAP: the alternative splicing annotation project. *Nucleic Acids Res*., **31**, 101–105.
16. Le,K., Mitsouras,K., Roy,M., Wang,Q., Xu,Q., Nelson,S.F. and Lee,C. (2004) Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data. *Nucleic Acids Res*., **32**, e180.
17. Xing,Y., Resch,A. and Lee,C. (2004) The multiassembly problem: reconstructing multiple transcript isoforms from EST fragment mixtures. *Genome Res*., **14**, 426–441.
18. Xing,Y., Xu,Q. and Lee,C. (2003) Widespread production of novel soluble protein isoforms by alternative splicing removal of transmembrane anchoring domains. *FEBS Lett*., **555**, 572–578.
19. Xu,Q. and Lee,C. (2003) Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. *Nucleic Acids Res*., **31**, 5635–5643.
20. Xu,Q., Modrek,B. and Lee,C. (2002) Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res*., **30**, 3754–3766.
21. Resch,A., Xing,Y., Alekseyenko,A., Modrek,B. and Lee,C. (2004) Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation. *Nucleic Acids Res*., **32**, 1261–1269.
22. Cusack,B.P. and Wolfe,K.H. (2005) Changes in alternative splicing of human and mouse genes are accompanied by faster evolution of constitutive exons. *Mol. Biol. Evol*., **22**, 2198–2208.
23. Lian,Y. and Garner,H.R. (2005) Evidence for the regulation of alternative splicing via complementary DNA sequence repeats. *Bioinformatics*, **21**, 1358–1364.
24. Roy,M., Xu,Q. and Lee,C. (2005) Evidence that public database records for many cancer-associated genes reflect a splice form found in tumors and lack normal splice forms. *Nucleic Acids Res*., **33**, 5026–5033.
25. Xing,Y. and Lee,C.J. (2005) Protein modularity of alternatively spliced exons is associated with tissue-specific regulation of alternative splicing. *PLoS Genet*., **1**, e34.
26. Chen,F.C., Wang,S.S., Chen,C.J., Li,W.H. and Chuang,T.J. (2006) Alternatively and constitutively spliced exons are subject to different evolutionary forces. *Mol. Biol. Evol*., **23**, 675–682.
27. Xing,Y., Wang,Q. and Lee,C. (2006) Evolutionary divergence of exon flanks: a dissection of mutability and selection. *Genetics*, **173**, 1787–1791.
28. Modrek,B. and Lee,C.J. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nature Genet*., **34**, 177–180.
29. Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med*., **75**, 694–698.
30. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*., **33**, D54–D58.
31. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*., **33**, D501–D504.
32. Birney,E., andrews,D., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cox,T., Cunningham,F., Curwen,V., Cutts,T. *et al*. (2006) Ensembl 2006. *Nucleic Acids Res*., **34**, D556–D561.
33. Blanchette,M., Kent,W.J., Riemer,C., Elnitski,L., Smit,A.F., Roskin,K.M., Baertsch,R., Rosenbloom,K., Clawson,H., Green,E.D. *et al*. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*., **14**, 708–715.
34. Hinrichs,A.S., Karolchik,D., Baertsch,R., Barber,G.P., Bejerano,G., Clawson,H., Diekhans,M., Furey,T.S., Harte,R.A., Hsu,F. *et al*. (2006) The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res*., **34**, D590–D598.