JuncDB: an exon–exon junction database

Michal Chorev[1,2], Lotem Guy[1,2] and Liran Carmel[1,*]

[1]Department of Genetics, The Alexander Silberman Institute of Life Sciences, Faculty of Science, The Hebrew University of Jerusalem, Edmond J. Safra Campus, Givat Ram, Jerusalem 91904, Israel and [2]The Rachel and Selim Benin School of Computer Science and Engineering, The Hebrew University of Jerusalem, Edmond J. Safra Campus, Jerusalem 91904, Israel

ABSTRACT

Intron positions upon the mRNA transcript are sometimes remarkably conserved even across distantly related eukaryotic species. This has made the comparison of intron–exon architectures across orthologous transcripts a very useful tool for studying various evolutionary processes. Moreover, the wide range of functions associated with introns may confer biological meaning to evolutionary changes in gene architectures. Yet, there is currently no database that offers such comparative information. Here, we present JuncDB (http://juncdb.carmelab. huji.ac.il/), an exon–exon junction database dedicated to the comparison of architectures between orthologous transcripts. It covers nearly 40 000 sets of orthologous transcripts spanning 88 eukaryotic species. JuncDB offers a user-friendly interface, access to detailed information, instructive graphical displays of the comparative data and easy ways to download data to a local computer. In addition, JuncDB allows the analysis to be carried out either on specific genes, or at a genome-wide level for any selected group of species.

INTRODUCTION

Spliceosomal introns are one of the main characteristics of eukaryotic species. All fully sequenced eukaryotic genomes have been shown to harbor at least a few introns, and many species—for example vertebrates—harbor on average multiple introns per gene (1). Although the evolutionary rate of intronic sequences typically resembles the rate of intergenic regions, introns are by no means function-less elements. In fact, many introns carry out diverse cellular functions and are involved in virtually all stages of mRNA processing, from transcription regulation through export and to quality control (2). This has rendered many introns essential in present day species, a property that is reflected in high conservation—not of their sequence but rather of their position within the coding exons (3–5).

The unique nature of introns, combined with their potential roles in eukaryotes biology, have triggered many comprehensive studies of their evolution, and revealed distinctive patterns of intron gain and loss throughout eukaryotic evolution (6–8). It is now well accepted that the past billion years of eukaryotic evolution consists primarily of intron loss, with a few notable episodes of intron gain in specific clades like metazoans and plants, and that the last eukaryotic common ancestor had been intron-rich, with about 50–70% of human intron density. These findings are compatible with the view of introns as slightly deleterious elements that invaded early eukaryotes and rapidly spread throughout their genomes, aided by the small effective populations of their hosts (9). Only later, some introns have gradually acquired function via opportunistic co-option by natural selection. Under this scenario introns that had acquired a function would be more resistant to loss, and will therefore show higher conservation of their position. Indeed, we have shown that the positions of introns that are functional due to miRNAs and snoRNAs embedded in their sequence, tend to be more conserved (5).

These considerations turned the identification of conserved intronic positions, and of intron gain and loss events, to a central step in studying evolution in eukaryotes. This is normally done by a comparative study of the exonic architecture of orthologous transcripts (6,7). However, there is currently no database that facilitates such comparative studies. Previous attempts to establish such a database, like ExoLocator (10), PIECE (11), Exon–Intron Database (12) and ExInt (13) lack in scope and available features, or are completely obsolete. Here, we present a new database dedicated to the comparative gene architecture of orthologous transcripts. JuncDB allows both large-scale cross-species comparative studies, as well as studies that focus on specific genes of interest.

MATERIALS AND METHODS

Data preparation

In a previous study, we compiled gene architecture data—namely, the positions of the exon–exon junctions—for 98 fully-sequenced eukaryotic species (5).

---

*To whom correspondence should be addressed. Tel: +972-2-6585103; Fax: +972-2-6584856; Email: liran.carmel@huji.ac.il

For the purpose of generating the database, missing data were filled-in and annotations were updated using Ensembl (14) and Ensembl Genomes (15) through the BioMart interface. Moreover, the data were extended to include 16 additional species, totaling in 114 fully-sequenced eukaryotes. Transcripts and proteins were divided to groups of orthologs using OrthoDB (16), OrthoMCL (17), Homologene (18) and Ensembl Compara (19). The latter was queried via its Perl API to detect orthologs based on human Ensembl transcript accessions. The sets of orthologs currently cover 88 species out of the 114 available. We used MUSCLE (20) to align the proteins within each group of orthologs, and then switched to the alignment at the mRNA level using Matlab's *seqinsertgaps* function. Finally, the positions of the exon–exon junctions, taken from the transcripts' annotations, were projected onto the multiple sequence alignment.

Database structure

JuncDB was designed using relational tables, and was implemented using Microsoft's SQL Server Management Studio 2008. Its website was built using ASP.NET MVC5 and Entity Framework 6 on Windows Server 2012 with IIS 8.5. Figure 1 shows the database's simplified diagram, which underlies the website design. The extended diagram, which was used during the database creation, is shown in Supplementary Figure S1. Each entity (transcript, protein, species, set of orthologs, orthology database, etc.) is represented by a table describing its main properties. For instance, 'JuncDB_OrthoGroups' describes sets of orthologs, and 'Orthologous_db' describes the orthology database from which the data were taken. Many-to-many relationships between entities are captured in auxiliary tables. For examples, each set of orthologs consists of multiple proteins/transcripts and each protein/transcript may belong to multiple sets of orthologs. These complex relationships are captured in the two tables 'JuncDB_OrthoGroups_has_Protein' and 'JuncDB_OrthoGroups_has_Transcript'.

The entity 'Transcript' is central to the database. It contains information on the lengths of the 3′ and 5′ untranslated regions (UTRs), and on the number of exons. Furthermore, it is connected in a many-to-many relationship to both the 'Intron' and the 'Exon' entities. 'Chromosome' has a field that holds the path to the chromosome sequence file, stored in UCSC 2bit format. Each 'Transcript' is also connected to a corresponding 'Protein' entity, and both are connected to their accession numbers tables. A description of the main database entities can be found in Supplementary Table S1. The relational tables can be freely downloaded from the database website (http://juncdb.carmelab.huji.ac.il/Download).

Phylogenetic tree

We generated the 88 species tree by combining topological data from NCBI (18), the Tree of Life Web Project (http://tolweb.org/tree/) and FlyBase (21). Divergence times were taken from TimeTree (22). As was already described (5), when divergence times were missing we used UPGMA

(23) to reconstruct the phylogeny of close species and estimated the divergence time using linear regression. The full species tree is shown in Figure 2 (generated using iTOL (24)). For each set of orthologs we generate a pruned version of the tree, containing only the species that have representatives in the set. This was done by extending our Comparative Genomics Toolbox for Matlab (http://carmelab.huji.ac.il/software/CG/cg.html).

RESULTS

Database organization

The most elementary building blocks of the database are exons and introns, which are the only entities that hold physical coordinates along the chromosome. The central entity of the JuncDB database is the transcript, made up from combinations of the elementary exons and introns. We focus on transcripts rather than on genes, because a gene may give rise to multiple transcripts, each having its own exonic architecture and evolutionary history. Different Splicing variants are considered different transcripts. The information held by the exons and the introns that make up a transcript, together with the information held by the transcript itself (like the lengths of the UTRs), provide a full virtualization of a genomic transcript (Supplementary Figure S1). While transcripts form the core of the database, we keep information on their associated genes, allowing for the user to easily work also at the gene level, if desired.

The relation of transcripts to proteins is straightforward. The transcript's virtualization allows an immediate reconstruction of its underlying mRNA sequence. A transcript is connected to an annotated protein if its mRNA is indeed translated to this protein.

Finally, the comparative aspect of the database comes about by grouping transcripts and proteins into sets of orthologs. Each set holds full information on the multiple sequence alignments of both the transcripts and the proteins that make it. If a gene has multiple splicing variants, the set of orthologs contains the single variant that most similar to the other transcripts in the set. Thus, different splicing variants of the same gene may be present in different sets of orthologs.

Data content

JuncDB contains the exonic architecture of 37 122 sets of orthologous transcripts (Table 1) across 88 eukaryotic species. Clearly, not all species have a representative in a set, and the average number of species per set is 14. These species span the major eukaryotic groups, and include representatives from metazoans, fungi, plants and many unicellular groups (Supplementary Table S2). In total, the sets of orthologs are made of 233 057 transcripts and proteins, which come from 42 347 different genes. Each transcript in the database is associated with its corresponding protein, its corresponding gene, the species in which it is found, all relevant accession numbers (Ensembl and RefSeq), and most importantly, the positions of the introns comprising it. Exon–exon junction positions are calculated on the gapped mRNA multiple sequence alignment.
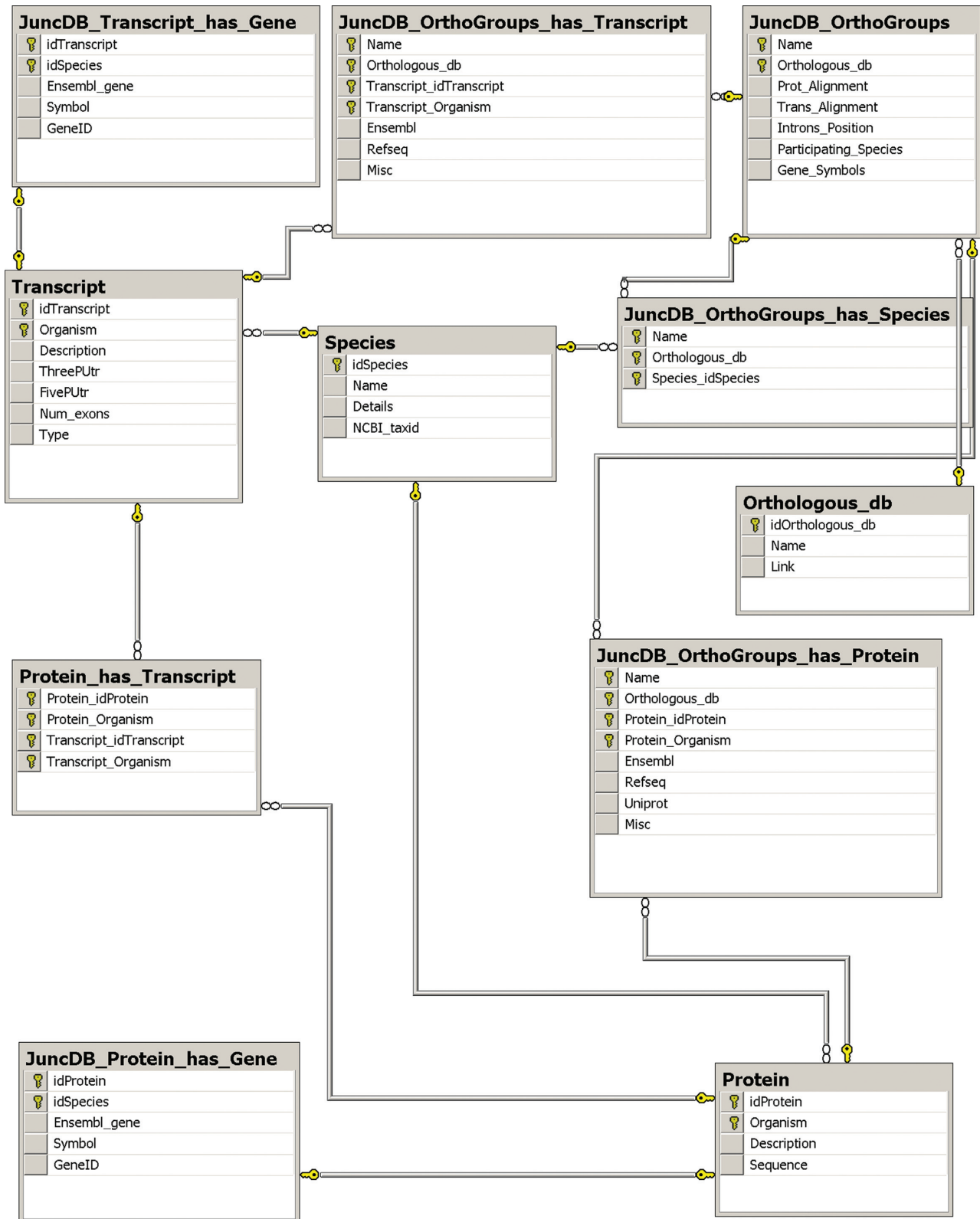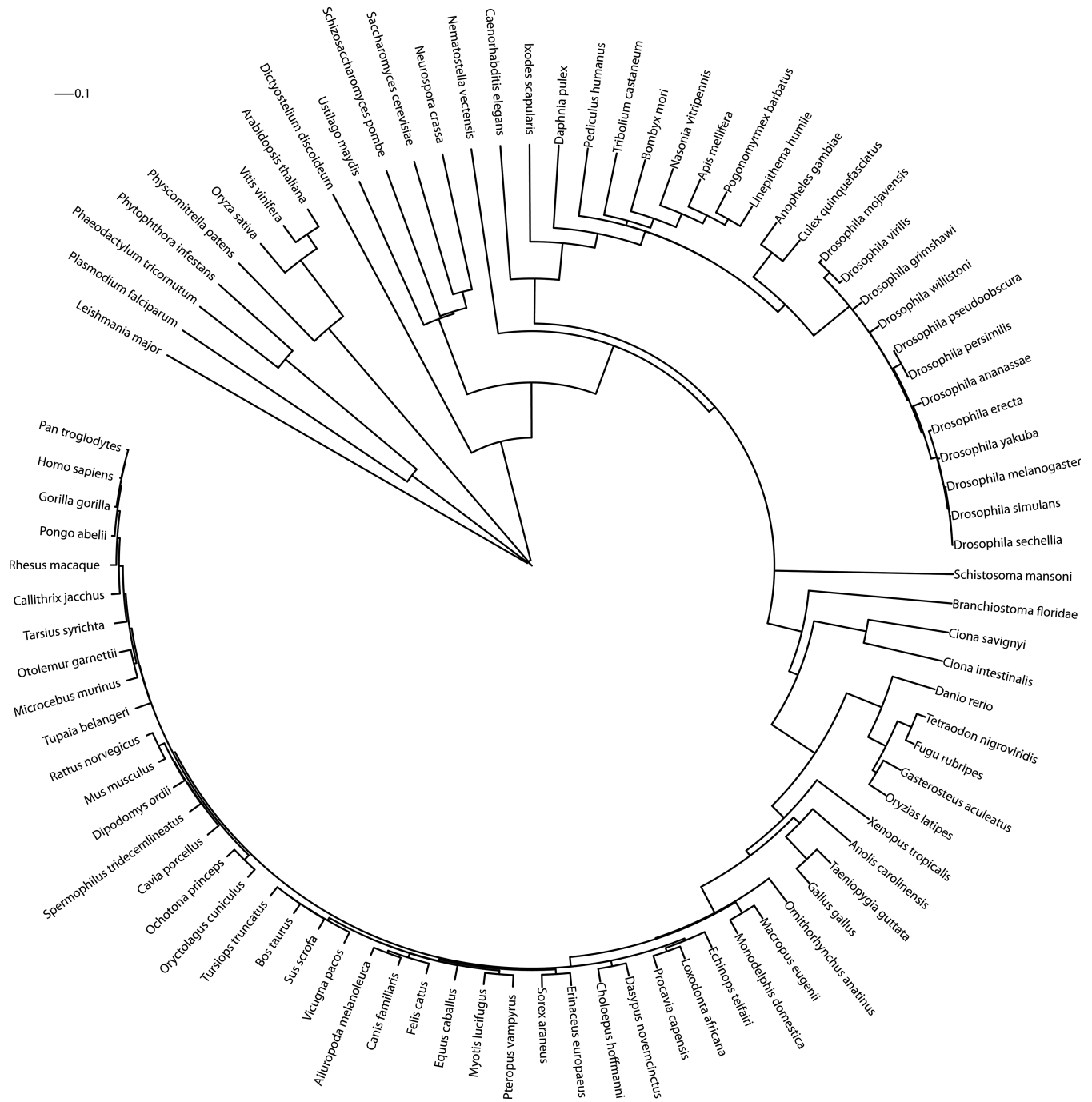
Figure 1. JuncDB relational database simplified diagram.

Figure 2. The phylogenetic tree of the 88 eukaryotic species in JuncDB.

Table 1. Orthology databases that are incorporated into JuncDB and number of sets of orthologs in each database

Ortholog database	# sets of orthologs
Compara	15 649
Homologene	620
OrthoDB	20 281
OrthoMCL	572
Total	37 122

Figure 3. JuncDB browse or query results page. (**A**) Download the current row (set of orthologs) information; (**B**) Download data on all sets of orthologs on the page; (**C**) Detailed description of the current row (set of orthologs).

JuncDB offers many tools that facilitate information extraction from the database. These include the phylogenetic tree connecting the species that make up each set of orthologs, all multiple alignments at the protein and the mRNA level, and both textual and graphical detailed description of the exonic architecture of the transcripts in each set of orthologs.

Website features

JuncDB (http://juncdb.carmelab.huji.ac.il/) provides a friendly interface to browse, query and visualize the exonic architecture of orthologous transcripts. The user can browse through the sets of orthologs, ordered by their name and by the orthology database from which they had been obtained (Figure 3). For each set, JuncDB displays its name ('orthologs set ID'), its source database ('orthology DB'), the symbol of the gene associated with it ('gene symbol') and the species that have a representative in it ('present in species'). Clicking the blue globe icon next to a gene symbol opens the UCSC Genome Browser (25) at the gene's location in the human genome. Clicking the 'Download' link (Figure 3A) downloads the full data of the set, including all text files and images. It is also possible to instantly download these data for all sets displayed in the current page by clicking the 'Download Page Data' link (Figure 3B).

Clicking on the 'Details' link (Figure 3C) displays all available data on the selected set of orthologs (Figure 4). At the very top, the page shows the orthologs set ID (Figure 4A), the orthology database (Figure 4B) and the gene symbol (Figure 4C). Gene symbols are not unique identifiers, and are determined by the following method: if there are human representatives in the set, show their gene symbols (could be more than one); otherwise, if other Ensembl species are part of the set, show the gene symbol held by the majority; otherwise, show the gene symbol held by the majority of all species. If there is a tie, multiple gene symbols are shown. Here too, the blue globe icons next to the gene symbols link each gene to its position in the human genome as it appears in the UCSC Genome Browser.

The next section in the 'Details' page shows the phylogenetic tree of the species in the set (Figure 4D), and the scheme of their mRNA multiple sequence alignment (Figure 4E). Each row in the alignment describes a single transcript in the set, and is labeled by the species name and the transcript ID in our database. Within each row vertical bars
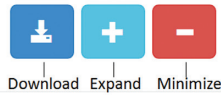
indicate exon–exon junctions, and so a rectangular block represents an exon. The number inside each rectangle indicates its length. Notably, these lengths are not necessarily the true exon lengths, but rather their lengths when projected on the multiple sequence alignment. If, in reality, a block consists of gaps, the actual length of the exon would be shorter than the length marked in the plot. Both the tree and the multiple alignment figures are vector images, allowing the user to zoom in and pan, as well as to hide the internal nodes of the tree or the exons lengths in the alignment.

The intron positions upon the mRNA multiple sequence alignment for each transcript in the set appear under the 'Intron Positions' section (Figure 5A). Similarly to the multiple sequence alignment scheme, each row corresponds to a single transcript in the set. Additional information on each of the transcripts is provided in the 'Transcripts' table (Figure 5B). This information may be used mainly to cross-link the transcript to external databases. Similar information on the corresponding proteins is given in the 'Proteins' table (Figure 5C). The last two sections provide the multiple sequence alignments themselves, at both the transcript (Figure 5D) and the protein (Figure 5E) levels. By clicking on the table name, the user can choose whether to display the information. Help is provided for each table through the orange information icon next to the link.

Querying the database

The database was built under the premise that it will be used for two main purposes: providing comparative information for a list of genes that are of particular interest to the researcher, or comparing gene architectures across a certain set of species. Accordingly, the database can be queried by using a list of gene/transcript/protein standard identifiers and/or by selecting a set of species. Currently, JuncDB supports the following identifiers: Ensembl transcript, Ensembl protein, Ensembl gene, RefSeq protein, RefSeq mRNA, Gene Symbol, Entrez Gene ID and set of orthologs name. The user can cross this list of identifiers with a list of species. By default, the query is then limited to sets of orthologs in which at least one of the selected species appears, as indicated by the radio button 'At least one of'. Alternatively, the user can choose to limit the results only to sets of orthologs in which all of the selected species appear, by clicking the 'All of' radio button (Figure 6). For instance, in order to get gene architecture data for all sets containing human, gorilla and orangutan, one should leave the list of
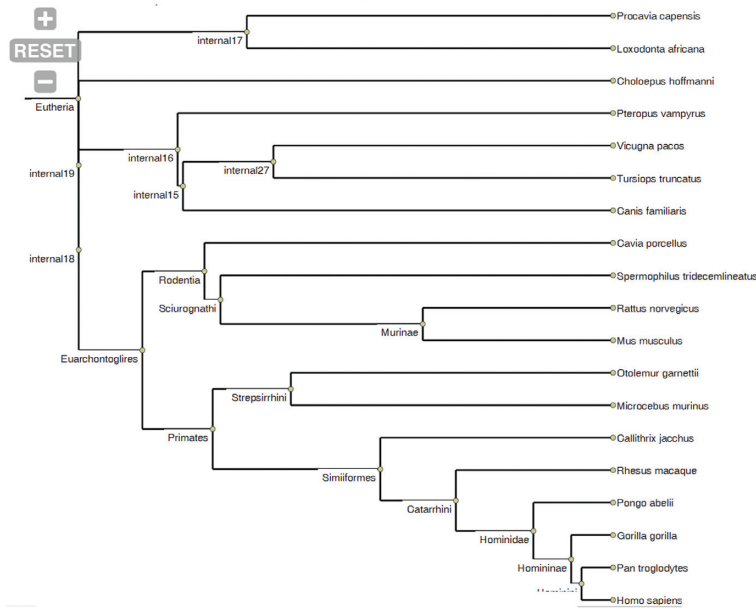
Figure 4. The upper part of JuncDB 'Details' page, providing information on a particular set of orthologs. (**A**) Orthologs set ID; (**B**) Source orthology database; (**C**) The gene symbol associated with the set; (**D**) The phylogenetic tree of the species represented in the set; (**E**) Scheme of the exonic architecture of all transcripts in the set. Each row describes a specific transcript, a vertical line is an exon–exon junction position and each block is an exon. The number within each block is the length of the exon in the multiple alignment.

ⓘ Intron Positions — A

Species	Transcript ID	Positions
Callithrix jacchus	11814	64
Canis familiaris	14473	64
Cavia porcellus	1148	64

ⓘ Transcripts — B

Species	ID	Accessions	Entrez Gene ID	Ensembl Gene
Callithrix jacchus	11814	ENSCJAT00000007569		ENSCJAG00000003951
Canis familiaris	14473	ENSCAFT00000038144	100683933	ENSCAFG00000024710
Cavia porcellus	1148	ENSCPOT00000026217	100715377	ENSCPOG0000002108

ⓘ Proteins — C

Species	ID	Accessions	Entrez Gene ID	Ensembl Gene
Callithrix jacchus	4607	ENSCJAP00000007163		ENSCJAG00000003
Canis familiaris	7937	ENSCAFP00000033820	100683933	ENSCAFG00000024
Cavia porcellus	869	ENSCPOP00000015716	100715377	ENSCPOG0000002

ⓘ Transcript Alignments — D

```
multiple sequence alignment in CLUSTALW format


Mus_musculus|13030          ------ATGGGGAGCCTACAGTTGACCCTCGTGCTCTTTGTCTTGCTC
Rattus_norvegicus|13121     ------ATGGGGAGCCTACAGTTGATCCTTGTGCTCTTTGTCTTGCTC
Cavia_porcellus|1148        TTAGTCATGAAGGGTCTACTCTTGGTCTTTGTGGTTCTTTTCTTGCTG
```

ⓘ Protein Alignments — E

```
multiple sequence alignment in CLUSTALW format


Mus_musculus|26089          --MGSLQLTLVLFVLLSYVPPVRSGVNMYIKRIYDTCWKLKGICF
Rattus_norvegicus|8410      --MGSLQLILVLFVLLSDVPPVRSGVNMYIRQIYDTCWKLKGHCF
Cavia_porcellus|869         LVMKGLLLVFVVLFLLSYVPPVRSGPNLYINRMFGSCWRMKGTCF
```

Figure 5. The lower part of JuncDB 'Details' page. (**A**) Intron positions projected upon the mRNA multiple alignment; (**B**) Information of the set's transcripts; (**C**) Information on the set's proteins; (**D**) The set's transcripts multiple sequence alignment; (**E**) The set's proteins multiple sequence alignment.

Species

○ At least one of   ○ All of

(hold the `ctrl`/`command` key for multiple selection)

```
--All--
Ailuropoda melanoleuca
Anolis carolinensis
Anopheles gambiae
Apis mellifera
Arabidopsis thaliana
Bombyx mori
Bos taurus
Branchiostoma floridae
Caenorhabditis elegans
Callithrix jacchus
Canis familiaris
Cavia porcellus
Choloepus hoffmanni
Ciona intestinalis
```
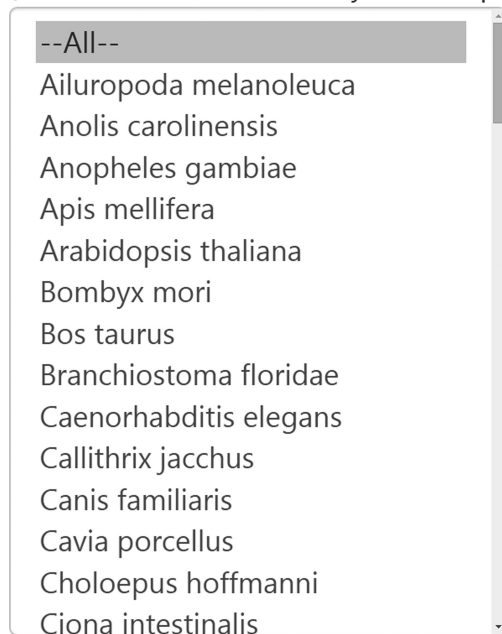
Figure 6. Filtering set of orthologs by species. The user can choose between 'At least one of', meaning it is sufficient for a set to contain at least one of the selected species, or 'All of', meaning the set must contain all selected species.

identifiers empty, check 'All of' under 'Species', and select 'Homo sapiens', 'Gorilla gorilla' and 'Pongo abelii' in the species list. The query page has a 'Load Example' button that exemplifies a typical database query. Detailed help is provided through the 'Need Help?' button.

Comparison to related databases

Very few databases provide comparative information on gene architectures. ExoLocator (10) puts up comparative analysis of protein-coding exons, but is restricted to Ensembl vertebrates and offers very limited options for comparative inference. PIECE (11) is a gene architecture database that provides both textual and graphical comparative information. However, it is restricted only to 25 species of plants. The Exon–Intron Database (12,26), last updated in 2006, provides exons and introns FASTA files for 12 species, but holds comparative information only among mammals. Information on gene architectures of 14 species is provided by ExonMine (27), but like the Exon-Intron database, it lacks comparative information. ExInt (13) provided gene architecture information for each GenBank entry, but is no longer reachable. In any case, the data was only available as flat files/MySQL dumps, and was missing a comparative tool and visualization capabilities.

Some tools enable gene architecture visualization, but lack the underlying database infrastructure. GenePainter (28) is an excellent visualization tool, but users cannot browse or query for specific genes. Rather, they must prepare in advance many input files that should be loaded to the website (protein multiple sequence alignment and gene architecture files for each species are the minimum). Moreover, in contrast to JuncDB, GenePainter does not allow genome-wide comparison by species. Namely, it is limited to plotting specific genes according to the user's interest, but cannot fetch comparative architectonic data for a set of species. Exalign (29), last updated in 2008, relies on multiple alignments of exon lengths, without consideration of the underlying sequences. Such alignments are expected to represent well changes in gene architecture for closely related species or for highly conserved proteins, but may miss more complex processes that lead to alterations in exon lengths. Moreover, it only allows pairwise comparisons between two known RefSeq transcripts, or a BLAST-like search against a fixed set of 14 species. GECA (30) was a visualization tool for gene exon/intron organization, but is no longer available.

CONCLUSIONS

JuncDB is a database for comparing exonic architectures within sets of orthologous transcripts, offering user-friendly web interfaces and extensive visualization tools. The database is very large, providing gene architecture information on 37 122 sets of orthologs covering 88 eukaryotic species. The database can be queried and browsed, and the results are given in both textual and graphical ways. Users can query the database by providing transcript/gene/protein IDs and/or by selecting a set of species. Thus, JuncDB makes it possible to investigate the gene architecture of specific genes or compare architectures across selected species in a genome-wide fashion. By matching the comparative gene architecture data to the phylogenetic species tree, JuncDB makes it possible to infer intron gain and loss events, and therefore also to study the potential functional roles of specific introns.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

REFERENCES

1. Rogozin,I.B., Carmel,L., Csuros,M. and Koonin,E.V. (2012) Origin and evolution of spliceosomal introns. *Biol. direct*, **7**, 11.
2. Chorev,M. and Carmel,L. (2012) The function of introns. *Front. Genet.*, **3**, 55.
3. Fedorov,A., Merican,A.F. and Gilbert,W. (2002) Large-scale comparison of intron positions among animal, plant, and fungal genes. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 16128–16133.

4. Rogozin,I.B., Wolf,Y.I., Sorokin,A.V., Mirkin,B.G. and Koonin,E.V. (2003) Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr. Biol.*, **13**, 1512–1517.

5. Chorev,M. and Carmel,L. (2013) Computational identification of functional introns: high positional conservation of introns that harbor RNA genes. *Nucleic Acids Res.*, **41**, 5604–5613.

6. Carmel,L., Rogozin,I.B., Wolf,Y.I. and Koonin,E.V. (2007) Patterns of intron gain and conservation in eukaryotic genes. *BMC Evol. Biol.*, **7**, 192.

7. Csuros,M., Rogozin,I.B. and Koonin,E.V. (2011) A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. *PLoS Comput. Biol.*, **7**, e1002150.

8. Roy,S.W. (2006) Intron-rich ancestors. *Trends Genet.*, **22**, 468–471.

9. Koonin,E.V. (2006) The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? *Biol. Direct*, **1**, 22.

10. Khoo,A.A., Ogrizek-Tomas,M., Bulovic,A., Korpar,M., Gurler,E., Slijepcevic,I., Sikic,M. and Mihalek,I. (2014) ExoLocator–an online view into genetic makeup of vertebrate proteins. *Nucleic Acids Res.*, **42**, D879–D881.

11. Wang,Y., You,F.M., Lazo,G.R., Luo,M.C., Thilmony,R., Gordon,S., Kianian,S.F. and Gu,Y.Q. (2013) PIECE: a database for plant gene structure comparison and evolution. *Nucleic Acids Res.*, **41**, D1159–D1166.

12. Saxonov,S., Daizadeh,I., Fedorov,A. and Gilbert,W. (2000) EID: the Exon-Intron Database-an exhaustive database of protein-coding intron-containing genes. *Nucleic Acids Res.*, **28**, 185–190.

13. Sakharkar,M., Long,M., Tan,T.W. and de Souza,S.J. (2000) ExInt: an Exon/Intron Database. *Nucleic Acids Res.*, **28**, 191–192.

14. Cunningham,F., Amode,M.R., Barrell,D., Beal,K., Billis,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fitzgerald,S. *et al.* (2015) Ensembl 2015. *Nucleic Acids Res.*, **43**, D662–D669.

15. Kersey,P.J., Allen,J.E., Christensen,M., Davis,P., Falin,L.J., Grabmueller,C., Hughes,D.S., Humphrey,J., Kerhornou,A., Khobova,J. *et al.* (2014) Ensembl Genomes 2013: scaling up access to genome-wide data. *Nucleic Acids Res.*, **42**, D546–D552.

16. Waterhouse,R.M., Tegenfeldt,F., Li,J., Zdobnov,E.M. and Kriventseva,E.V. (2013) OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res.*, **41**, D358–D365.

17. Fischer,S., Brunk,B.P., Chen,F., Gao,X., Harb,O.S., Iodice,J.B., Shanmugam,D., Roos,D.S. and Stoeckert,C.J. Jr. (2011) Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. *Curr. Protoc. Bioinformatics*, Chapter 6, Unit 6, 12, 1–19.

18. NCBI Resource Coordinators (2015) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **43**, D6–D17.

19. Vilella,A.J., Severin,J., Ureta-Vidal,A., Heng,L., Durbin,R. and Birney,E. (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, **19**, 327–335.

20. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

21. Drysdale,R. and FlyBase,C. (2008) FlyBase : a database for the Drosophila research community. *Methods Mol. Biol.*, **420**, 45–59.

22. Hedges,S.B., Dudley,J. and Kumar,S. (2006) TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics*, **22**, 2971–2972.

23. Sokal,R.R. and Michener,C.D. (1958) A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull.*, 1409–1438.

24. Letunic,I. and Bork,P. (2011) Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.*, **39**, W475–W478.

25. Karolchik,D., Barber,G.P., Casper,J., Clawson,H., Cline,M.S., Diekhans,M., Dreszer,T.R., Fujita,P.A., Guruvadoo,L., Haeussler,M. *et al.* (2014) The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.*, **42**, D764–D770.

26. Shepelev,V. and Fedorov,A. (2006) Advances in the Exon-Intron Database (EID). *Brief. Bioinform.*, **7**, 178–185.

27. Mollet,I.G., Ben-Dov,C., Felicio-Silva,D., Grosso,A.R., Eleuterio,P., Alves,R., Staller,R., Silva,T.S. and Carmo-Fonseca,M. (2010) Unconstrained mining of transcript data reveals increased alternative splicing complexity in the human transcriptome. *Nucleic Acids Res.*, **38**, 4740–4754.

28. Muhlhausen,S., Hellkamp,M. and Kollmar,M. (2015) GenePainter v. 2.0 resolves the taxonomic distribution of intron positions. *Bioinformatics*, **31**, 1302–1304.

29. Pavesi,G., Zambelli,F., Caggese,C. and Pesole,G. (2008) Exalign: a new method for comparative analysis of exon-intron gene structures. *Nucleic Acids Res.*, **36**, e47.

30. Fawal,N., Savelli,B., Dunand,C. and Mathe,C. (2012) GECA: a fast tool for gene evolution and conservation analysis in eukaryotic protein families. *Bioinformatics*, **28**, 1398–1399.