White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Leeds.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/147319/

Version: Accepted Version

Article:
Wang, D (2019) IntronDB: a database for eukaryotic intron features. Bioinformatics, 35 (21). btz242. pp. 4400-4401. ISSN: 1367-4803

https://doi.org/10.1093/bioinformatics/btz242

© The Author(s) 2019. Published by Oxford University Press. All rights reserved. This is a pre-copyedited, author-produced version of an article accepted for publication in Bioinformatics following peer review. The version of record Wang, D (2019) IntronDB: a database for eukaryotic intron features. Bioinformatics. btz242. ISSN 1367-4803 is available online at: https://doi.org/10.1093/bioinformatics/btz242

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Applications Note

Genome analysis

IntronDB: a database for eukaryotic intron features

Dapeng Wang[1,2,*]

[1]Department of Plant Sciences, University of Oxford, Oxford OX1 3RB, UK and [2]LeedsOmics, University of Leeds, Leeds LS2 9JT, UK

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: The rate and extent of unbalanced eukaryotic intron changes exhibit dynamic patterns for different lineages of species or certain functional groups of genes with varied spatio-temporal expression modes affected by selective pressure. To date, only a few key conserved splicing signals or regulatory elements have been identified in introns and little is known about the remaining intronic regions. To trace the evolutionary trajectory of spliceosomal introns from available genomes under a unified framework, we present IntronDB, which catalogues ~50000000 introns from over 1000 genomes spanning the major eukaryotic clades in the tree of life. Based on the position of introns relative to coding regions, it categorizes introns into three groups such as 5'UTR, CDS and 3'UTR and subsequently divides CDS introns into three categories such as phase 0, phase 1 and phase 2. It provides the quality evaluation for each sequence entry and characterizes the intronic parameters including number, size, sequence composition and positioning information as well as the features for exons and genes, making possible the comparisons between introns and exons. It reports the dinucleotides around the intron boundary and displays the consensus sequence features for all introns, small introns and large introns for each genome. By incorporating the taxonomic assignment of genomes, it performs high-level or genome-wide statistical analysis for single feature and coupled features both in a single genome and across multiple genomes. It offers functionalities to browse the data from representative protein-coding transcripts and download the data from all transcripts from protein-coding genes.
Availability: http://www.nextgenbioinformatics.org/IntronDB
Contact: dapeng.wang@plants.ox.ac.uk
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Exon-intron junction structures emerge in eukaryotic genomes as a consequence of complicated evolutionary history in terms of genome architecture and transcription machinery (Deutsch and Long, 1999). They provide diversified choices for the genomes in the process of removing the introns and generating distinctive types of transcripts by shifting from constitutive splicing to alternative splicing at the cost of considerable energy and time (Chorev and Carmel, 2012). It is documented that distinct groups of lineages have developed obviously different strategies for the splicing of large introns in order to complete the splicing events efficiently (Shepard, et al., 2009). To address the challenge posed by systematic comparisons among a broad range of species, we decided to build IntronDB to help the researchers to examine the hypotheses that are relevant to spliceosomal introns for the genomes that are publicly available. We calculated the fundamental intron-relevant genomic features and utilized a variety of graphical techniques to reveal the nature of those features by considering the distributions of single variable and two variables as well as those examined under the framework of taxonomical parent-child information.

2 Methods

We retrieved the DNA sequences and gene annotation sets (GTF files) from Ensembl databases (https://www.ensembl.org, http://www.ensemblgenomes.org) and parsed the information and extracted the sequences of introns, exons and transcripts based on the two types of datasets. The core data of this database are only derived from the datasets where each

protein-coding gene is represented by its longest protein-coding transcript. To cope with the frameshift introns, we only kept introns with size >=10 bp in the downstream analysis. For the sake of quality control, we evaluated the fraction of the equivocal bases or N nucleotides in independent introns and exons as well as in the whole genes in the five categories such as 0%, 0%-1%, 1%-10%, 10%-50% and >50%. In addition, we estimated the genome-wide N% for introns, exons, genes and genomes to give an overview of the sequencing, assembly and annotation qualities of each genome. We defined the first two and the last two nucleotides of each intron as donor- and acceptor-splice pairs and computed the proportion of introns with canonical pairs (GT-AG) and non-canonical pairs (e.g., AT-AC) for a given genome. To measure the sequence conservation scores, we collected the data from the UCSC Genome Browser (https://genome.ucsc.edu) and took the average values of all available phastCons and phyloP scores for each site within given introns and exons for five model organisms. We analyzed the guanine-cytosine content (GC content) and pyrimidine content for all elements and only the unambiguous portion of sequences was taken into account during the calculation. To scale some features to their genomic baselines, we computed the standard scores of size, GC content and pyrimidine content for introns and exons. Based on the observation of the distributions of intron sizes in a few representative genomes, we classified all introns into two groups based on their sizes such as small introns (<=1000 bp) and large introns (>1000 bp). In particular, we examined two categories of introns with extreme sizes including super small introns (<=100 bp) and super large introns (>50000 bp) in order to discover the genomes with peculiar intron sizes. We divided all introns into three groups such as 5'UTR, CDS and 3'UTR according to the relative position of intron to the coding sequence in the transcript and for CDS introns, we further categorized them into three classes including phase 0, phase 1 and phase 2 depending on the position of intron in the three codon positions. To associate each intron with its host gene, we labelled genes based on the size and number of the introns that the gene carries such as intronless genes, single-intron genes, super-small-intron-containing genes, super-small-intron-only genes, small-intron-containing genes, small-intron-only genes, large-intron-containing genes, large-intron-only genes, super-large-intron-containing genes and super-large-intron-only genes. We used "Intron Index Position" and "Intron Base Position" to characterize the relative position of each intron along the direction of transcription when the rank and distance were considered. In order to generate the sequence logos for introns surrounding the splice sites, we employed WebLogo 3.6.0 to draw the sequence features of four kinds of nucleotides for all introns (50 bp), small introns (30 bp) and large introns (100 bp) separately with regards to the donor splice sites and acceptor splice sites (Crooks, et al., 2004). Based on the reasonable criterion, we produced images for all introns from genomes that have at least 100 introns with size >=50 bp, and used the top 1000 introns (the smallest 1000 introns or the largest 1000 introns) for small introns or large introns from genomes that have >=2000 introns with size >=30 bp and harbor >=1000 introns with size >=100 bp. Moreover, all sequences with ambiguous nucleotides were filtered out before the production of the images. To facilitate the comparison among genomes, all organisms were assigned the taxonomy IDs and names based on the taxa annotation data from the NCBI taxonomy database (https://www.ncbi.nlm.nih.gov/taxonomy).

3 Results

The current release of IntronDB is composed of 5 sub-databases that correspond to the respective original Ensemble resources (Supplementary Fig. S1). For each sub-database, the entries have been organized into four groups such as genome, gene, exon and intron and for each group we loaded one text file into a table in the back-end MySQL database so that all the data are viewable from the front-end browser in a timely manner. For instance, the genome section provides a general estimate of all the important attributes based on the primary definitions and advanced gene categorizations over a genome scale as well as the statistical measurements of the attributes from each intron, exon or gene such as mean, standard deviation, minimum, the first quartile, median, the third quartile and maximum. Besides, the "compare" module provides a way to compare the features between two genomes. For intron, exon and gene section, the browsers have two layers in the structure where the first layer displays the selected brief information about each entry in a page such as identifier, location, strand, start, end, GC content, and transcript id, which allows to compare the properties of different entries at a time, and the second layer shows all the calculated values for each specific entry. It is worth noting that for intron section the fields include source, species name, intron ID, location, strand, intron start, intron end, intron size, intron GC (%), intron pyrimidine (%), intron N (%), intron donor-acceptor splice pair, intron type, intron phase, intron index position, intron base position, transcript ID, intron number and intron rank. The highlight of the functionalities is the "gallery" section, which provides five groups of graphical presentations within a rational layout (Supplementary Fig. S2-S11). The first group explores the statistical properties of basic genomic parameters using a wide range of chart types such as density plot, barplot, boxplot and pie chart. The second group produces the scatter plots of two genomic variables with the binning techniques in order to reveal their potentially complex relationships or correlations. The third group investigates the scatter plots of compared genome-wide features among a number of genomes from an evolutionary perspective through the introduction of taxonomic ranks of each organism. The fourth group compares mean against standard deviation of all available attributes across a group of genomes. The fifth group illustrates the logo of sequence features within the intronic regions surrounding the exon-intron junction based on the multiple alignments of nucleotides, enabling the discovery of the significant splicing signals and the comparison between small introns and large introns in a genome in terms of regulatory elements. The representative protein-coding datasets for genomes, transcripts, exons and introns alongside with the flat files containing information about all transcripts without any filtration for transcripts, exons and introns are downloadable in the download section. In the next release, our plans include enriching the classes of introns by integrating the intronic data from non-coding RNAs and providing a more flexible and customized interface for users to perform the analysis and visualization of user-defined group of attributes.

Conflict of Interest: none declared.

References

Chorev, M. and Carmel, L. (2012) The function of introns, *Front Genet*, **3**, 55.

Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) WebLogo: a sequence logo generator, *Genome Res*, **14**, 1188-1190.

Deutsch, M. and Long, M. (1999) Intron-exon structures of eukaryotic model organisms, *Nucleic Acids Res*, **27**, 3219-3228.

Shepard, S., McCreary, M. and Fedorov, A. (2009) The peculiarities of large intron splicing in animals, *PLoS One*, **4**, e7853.