

Release Notes

ASD - Release 3

ATD - Release 2

Mouse - Feb 2006

Human - Dec 2005

The release has been generated by:

Eleanor Whitfield - Co-ordinator

Gautier Koscielny - Technical lead

Vincent Le-Textier - Software engineer

Vasudev Kumanduri - Software engineer

Chellappa Gopalakrishnan - Software engineer

Table of contents:

Introduction

Access

Changes in current release

Forthcoming changes

Statistics

Contact details

Citation

Acknowledgements

Introduction:

The Alternative Splicing Database and Alternative transcript Database projects are creating a database of alternative splice events and transcripts of genes from human and mouse. Full length transcripts are generated with the aim of understanding the mechanism of alternative splicing on a genome-wide scale.

The current release of the human genome consists of:

16,599 genes, 15,608 have more than one splice isoform, with an average of 5 splice patterns per gene.

15,746 transcripts are annotated as full length with a transcription start site and a poly(A).

The current release of the mouse genome consists of:

16,020 genes, 12,619 have more than one splice isoform, with an average of 2.8 splice patterns per gene.

1,400 transcripts are annotated as full length with a transcription start site and a poly(A).

More comprehensive statistics are available at the end of this document.

The data in this release is generated for genes from Ensembl version 36.35i for human and version 37.34e for mouse. The extracted nucleotide region includes the gene region as defined in Ensembl with an extension at the 5' and 3' ends by 3000 bases for human and 10,000 bases for mouse. Human EST and mRNA (transcript) sequences are mapped to these extended gene regions. Transcript confirmed introns and exons are delineated from these alignments. The matching transcript sequences are further classified into groups, each group represents an isoform splice pattern. Each group is represented by a transcript representative structure (as defined by having the most introns). Isoform peptide translations are also presented.

Isoform splice patterns are compared with one another to delineate the alternative events. The basic events that are identified in this work are: exon isoforms (extension/truncation of an exon), intron isoforms (extension/truncation of an intron), cassette exons (an exon is present in one transcript but absent in an isoform of the transcript), mutually exclusive exons (exons are used in alternative transcripts in a mutually exclusive manner) and intron retention (a nucleotide region is used as an exon in a transcript while it is an intron in an alternative transcript). The latter three events (namely cassette exon, mutually exclusive exon and intron retention) are further characterised as 'complex' or 'simple' depending on whether the 5' or/and 3' flanking exons also undergo modifications (e.g. the flanking exon may be extended or truncated or the exon that flanks a retained intron is a cassette or mutually exclusive event). Introns/exons are annotated for splice signals such as donor/acceptor sites, branch points, and polypyrimidine tracts. Conserved exons/introns/events in the orthologous genes from human and mouse have been identified and are annotated in the database. SNP positions and alleles used have been mapped to our data and we display them for isoform splice patterns as well as for individual events. Annotation pertaining the expression states of the isoforms is being added to the data. Subtractive library expression queries can now be raised from the interfaces

Each transcript is scrutinised for the presence of a poly(A) tail, poly(A) sites upstream of the cleavage site and for a transcription start site (TSS).

The manually annotated database, AEdb, has been integrated to some extent with AltSplice (for both human and mouse entries). Entries that are common between AltSplice and AEdb are associated and are indicated so in the display pages that is resultant of queries to AltSplice and/or AEdb. Queries can be raised for common entries. AltSplice exons and splice events that have experimental evidence from AEdb are indicated so. In addition, we have built a wrapper that passes on queries to both the AEdb and AltSplice.

Access:

Access to the data from the automatic pipeline is via the Simple all text query on the

home pages:

<http://www.ebi.ac.uk/asd/>

<http://www.ebi.ac.uk/atd/>

An advanced search is available:

<http://www.ebi.ac.uk/asd-srv/Index.cgi?method=MAIN&product=WRAPPER&visit=first>

<http://www.ebi.ac.uk/asd-srv/Atd.cgi?method=MAIN&product=WRAPPER&visit=first>

Download of associated flat files is available:

<http://www.ebi.ac.uk/asd/altsplice/index.html>

<http://www.ebi.ac.uk/atd/download.html>

Changes in current release:

1)

Transcript structure has been augmented by the prediction of Transcription Start Sites (TSS). This is achieved by aligning 5' oligo capped mRNA to the predicted transcript. For human, the 1.3 million 5' end sequences (Ref. Kimura et al., 2006) are blasted against the EST/mRNA that have been found to confirm transcript structures. For mouse, 60,770 full-length cDNAs have been blasted (Ref. Okazaki et al., 2002). The alignments and subsequent mapping automatically link the TSS sites to a transcript.

In the post-processing steps we apply stringent conditions to select the meaningful high scoring pairs:

a) The first matching base must be base number 1 of the EST/mRNA.

b) The start position of the transcript sequence must be upstream of the first exon.

2)

Flanking untranslated DNA for the mouse gene models have been extended from 3kb to 10kb. Both human and mouse will be updated for the next release in early 2007.

3)

AltExtron has been removed from the interfaces. This database is no longer maintained at the EBI. The database can be found at:

<http://bit.uq.edu.au/altExtron/>

4)

Sequence ontology terms are now used to describe transcript structure.

<http://www.sequenceontology.org/>

5)

The exon index that is generated as a download file, is now also available in GFF3 format.

Unique identifiers have been assigned to the exons with the format EXONnnnnnnnnnn.

These unique identifiers will be mapped between releases.

Forthcoming changes:

1)

Renaming the database:

The ASD and ATD databases will be combined into one for the next release early in 2007. The new database will be named ASTD - Alternative splicing and transcript database, and will exhibit no loss of functionality or data.

The new database release will be accompanied with greatly improved web interfaces.

Snapshots of the new interfaces can be seen here.

Please feel free to make any comments concerning these improvements.

2)

Plans for 2007 include addition of a third species, possibly rat.

3)

Unique identifiers are to be assigned to:

Gene: ENSG

Exons: EXON

Introns: INTR

Splice event: EVEN

Transcript: TRAN

Translation: PEPT

Transcription start site: ATSS

Poly(A): POLY

4)

The download files will be available in different formats including chaos-xml and gff3.

Comprehensive Statistics:

ASD : <http://www.ebi.ac.uk/asd/altsplice/Statistics.html>

ATD : <http://www.ebi.ac.uk/atd/statistics.htm>

For all queries please contact:

ASTD team

The EMBL Outstation - The European Bioinformatics Institute

Wellcome Trust Genome Campus

Hinxton
Cambridge CB10 1SD
United Kingdom

Telephone: (+44 1223) 494 680
Telefax: (+44 1223) 494 468

We welcome any comments/questions about the data.
Please go to <http://www.ebi.ac.uk/support/index.php?query=ASD/ATD> with any queries.

Citation:

If you want to cite ASD in a publication, please use one of the following reference:
Stamm S, Riethoven J-JM, Le Texier V, Gopalakrishnan C, Kumanduri V, Tang Y, Barbosa-Morais NL, Thanaraj TA. ASD: a bioinformatics resource on alternative splicing. *Nucleic Acids Res* 2006 34: D46-D55.

If you want to cite ATD in a publication, please use one of the following reference:
Le Texier, V., Riethoven, J-J., Kumanduri, V., Gopalakrishnan, C., Lopez, F., Gautheret, D. and Thanaraj, T.A. (2006) AltTrans: Transcript pattern variants annotated for both alternative splicing and alternative polyadenylation. *BMC Bioinformatics* 7: 169 (2006).

Acknowledgements:

ASD consortium members:

Stefan Stamm, Institute of Biochemistry, University Erlangen Nuremberg, 91054 Erlangen, GERMANY.

Peer Bork, Structural and Comp. Biology Dept., European Molecular Biological Laboratory, 69177 Heidelberg, GERMANY.

Roderic Guigo, Genomics Laboratory Group, IMIM/Research on Biomedical Informatics, 08003, Barcelona, SPAIN.

Laurant Bracco, Exonhit Therapeutics, 75013 Paris, FRANCE.

Hermona Soreq, Department of Biological Chemistry, The Hebrew University of Jerusalem, 91904 Jerusalem, ISRAEL.

Rolf Apweiler, EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD,

Juan Valcarcel, Gene Expression Programme, European Molecular Biological Laboratory, 69177 Heidelberg, GERMANY.

ATD consortium members:

Peer Bork, Structural and Comp. Biology Dept., European Molecular Biological Laboratory, 69177 Heidelberg, GERMANY.

Christiane Dascher-Nadel, INSERM Transfert, France

Daniel Gautheret, INSERM, France

Roderic Guigo, Genomics Laboratory Group, IMIM/Research on Biomedical Informatics, 08003, Barcelona, SPAIN.

Winston Hide, SANBI, South Africa

Magnus von Knebel, University of Heidelberg, Germany

Jans Reich, Max-Delbruck-Centrum fur Molecular Medizin, Germany

Rolf Apweiler, EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD,

Jaak Vilo, Estonian Biocenter, Estonia

Eurasnet consortium members:

Alberto R. Kornblihtt, FCEN-UBA Buenos Aires Argentina

Andrea Barta, MUW Vienna Austria

Jorgen Kjems, UAAR Århus Denmark

Christiane Branlant, CNRS Nancy France

Jamal Tazi, CNRS Montpellier France

James Stévenin, IGBMC-GIE Illkirch France

Bertrand Séraphin, CNRS Gif sur Yvette France

Reinhard Lührmann, MPG Göttingen Germany

Stefan Stamm, UERLN Erlangen Germany

Albrecht Bindereif, LUG Giessen Germany

Peer Bork, EMBL Heidelberg Germany

Karla Neugebauer, MPG Dresden Germany

Gil Ast, TAU Tel Aviv Israel

Hermona Soreq, HUJI Jerusalem Israel

Francisco Baralle, ICGEB Trieste Italy

Giuseppe Biamonti, CNR Pavia Italy

Glauco Tocchini-Valentini, CNR Monterotondo Scalo Italy

Artur Jarmolowski, AMU Poznan Poland

Maria Carmo-Fonseca, IMM Lisbon Portugal

Juan Valcárcel, CRG Barcelona Spain

Göran Akusjärvi, UU Uppsala Sweden

Angela Krämer, UNIGE Geneva Switzerland

Daniel Schümperli, UNIBE Bern Switzerland

Rolf Apweiler, EMBL Hinxton UK

Jean Beggs, UEDIN Edinburgh UK

John Brown, SCRI Dundee UK

Javier F. Caceres, MRC Edinburgh UK

Ian Eperon, Unileic Leicester UK

Angus Lamond, UNIVDUN Dundee UK
Chris Smith, UCAM-DBIOC Cambridge UK

References

Kouichi Kimura et al. 'Diversification of transcriptional modulation: Large-scale identification and characterization of putative alternative promoters of human genes.', *Genome Research* 16:55-65, 2006

Okazaki et al., 'Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs', *Nature* 420, 563-573, 2002