Genomics 93 (2009) 213–220

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

ASTD: The Alternative Splicing and Transcript Diversity database

Gautier Koscielny [a], Vincent Le Texier [a], Chellappa Gopalakrishnan [a], Vasudev Kumanduri [a], Jean-Jack Riethoven [a,1], Francesco Nardone [a,2], Eleanor Stanley [a], Christine Fallsehr [b], Oliver Hofmann [c], Meelis Kull [d,e], Eoghan Harrington [f], Stéphanie Boué [f], Eduardo Eyras [g], Mireya Plass [g], Fabrice Lopez [h], William Ritchie [h], Virginie Moucadel [h,3], Takeshi Ara [h,4], Heike Pospisil [i], Alexander Herrmann [j], Jens G. Reich [k], Roderic Guigó [g,l], Peer Bork [f], Magnus von Knebel Doeberitz [b], Jaak Vilo [d,e], Winston Hide [c], Rolf Apweiler [a], Thangavel Alphonse Thanaraj [a,5], Daniel Gautheret [h,*,6]

[a] European Bioinformatics Institute, European Molecular Biology Laboratory, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK
[b] Department of Applied Tumor Biology, Institute of Pathology, University of Heidelberg, Im Neuenheimer Feld 220, D-69120 Heidelberg, Germany
[c] South African National Bioinformatics Institute, University Western Cape, Private Bag X17, Bellville 7535, South Africa
[d] Estonian Biocenter, Riia 23b, 51010 Tartu, Estonia
[e] University of Tartu, Liivi 2, 50409 Tartu, Estonia
[f] Structural and Computational Biology Unit, European Molecular Biology Laboratory, Meyerhofstraβe 1, 69117 Heidelberg, Germany
[g] Grup de Recerca en Informatica Biomedica, Institut Municipal d'Investigacio Medica, Universistat Pompeu Fabra, Plaça de la Mercè, 10-12, 08002 Barcelona, Spain
[h] INSERM U928, Université de la Méditerranée, case 928 - 163, Avenue de Luminy, 13288 Marseille cedex 09, France
[i] Center for Bioinformatics, University of Hamburg, Bundesstr.43, 20146 Hamburg, Germany
[j] Institute for Clinical Molecular Biology, Christian-Albrechts University Kiel, University Hospital Schleswig-Holstein, 24105 Kiel, Germany
[k] Dept. of Bioinformatics, Max-Delbrueck-Center for Molecular Medicine Berlin-Buch, Robert-Roessle-Str. 10, 13125 Berlin, Germany
[l] Center for Genomic Regulation, C/ Dr. Aiguader, 88 - 08003 Barcelona, Spain

ARTICLE INFO

Article history:
Received 24 April 2008
Accepted 5 November 2008
Available online 24 December 2008

Keywords:
Alternative transcription
Alternative splicing
Alternative polyadenylation
Alternative initiation

ABSTRACT

The Alternative Splicing and Transcript Diversity database (ASTD) gives access to a vast collection of alternative transcripts that integrate transcription initiation, polyadenylation and splicing variant data. Alternative transcripts are derived from the mapping of transcribed sequences to the complete human, mouse and rat genomes using an extension of the computational pipeline developed for the ASD (Alternative Splicing Database) and ATD (Alternative Transcript Diversity) databases, which are now superseded by ASTD. For the human genome, ASTD identifies splicing variants, transcription initiation variants and polyadenylation variants in 68%, 68% and 62% of the gene set, respectively, consistent with current estimates for transcription variation. Users can access ASTD through a variety of browsing and query tools, including expression state-based queries for the identification of tissue-specific isoforms. Participating laboratories have experimentally validated a subset of ASTD-predicted alternative splice forms and alternative polyadenylation forms that were not previously reported. The ASTD database can be accessed at http://www.ebi.ac.uk/astd.

Introduction

Transcript expression in eukaryotes is subject to variation at three main biological stages: transcription initiation, splicing and polyadenylation. In mammals, most genes undergo some kind of alternative transcription. Current data for human indicates that at least 81% of genes are subject to alternative transcription initiation [1], 69% to alternative splicing [2] and 60% to alternative polyadenylation [3]. Abnormal expression of alternative transcripts has been linked to multiple diseases, especially to cancer [4]. The sheer number and wide biological impact of alternative transcripts (ATs) has created a high demand for tools enabling the identification, classification, functional annotation and expression profiling of ATs in the genomes of major model organisms. To meet this demand, several AT databases have been developed based on large-scale mappings or assemblies of

* Corresponding author. Fax: +33 1 69 15 46 29.
E-mail address: daniel.gautheret@u-psud.fr (D. Gautheret).
[1] Current address: Center for Biotechnology, University of Nebraska-Lincoln, 1901 Vine Street, Lincoln, NE 68588, USA.
[2] Current address: WebOn ASm Postboks 2198, 3103 Tønsberg, Norway.
[3] Current address: Laboratoire de Transduction du Signal, Institut de Recherche en Technologies et Sciences pour le Vivant – CEA, 17 Rue des Martyrs, 38054 Grenoble, France.
[4] Current address: Kazusa DNA Research Institute, 2-6-7 Kazusa-kamatari, Kisarazu, Chiba 292-0818, Japan.
[5] Current address: CRS4-Bioinformatica, Parco Scientifico e Technologico, POLARIS, Edificio 3, 09010 Pula (CA), Sardinia, Italy.
[6] Current address: Univ. Paris-Sud 11, CNRS, UMR8621, Bat 400, 91405 Orsay cedex, France.

0888-7543/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2008.11.003

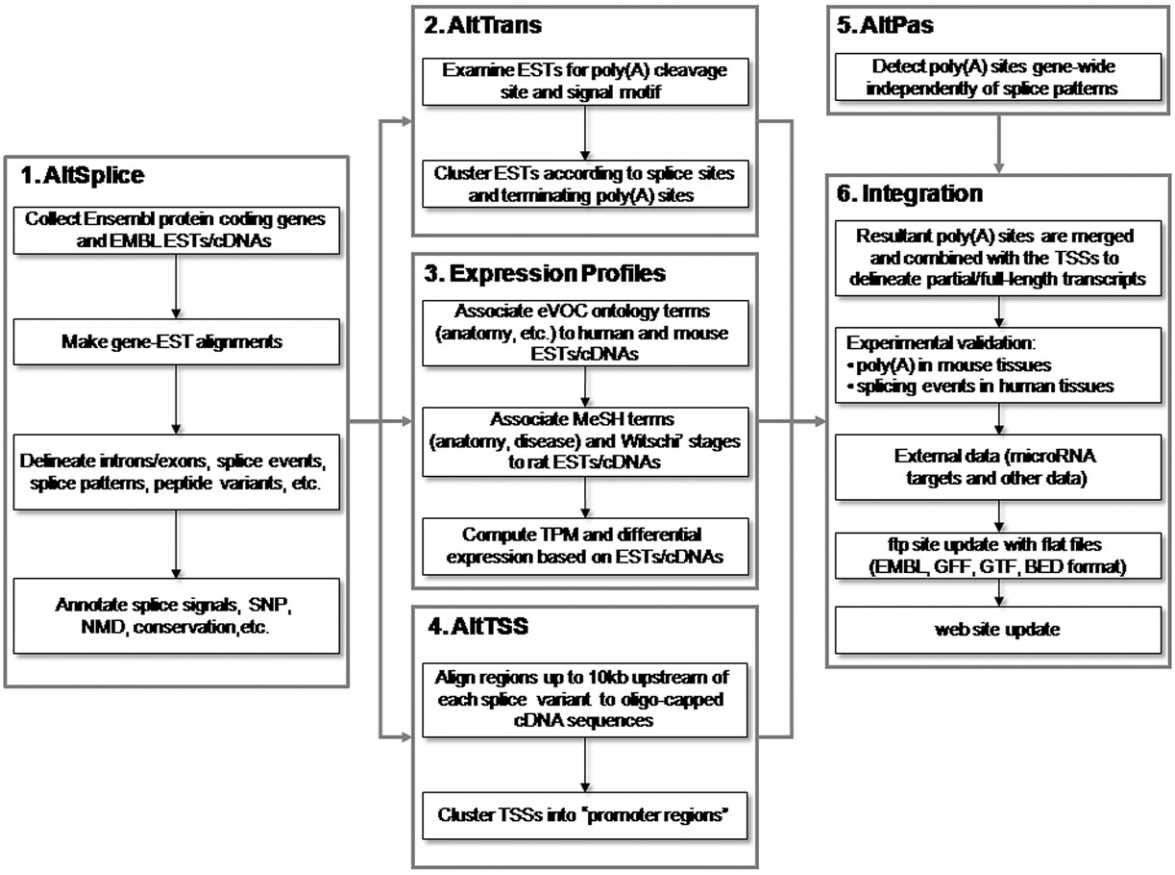


Fig. 1. Flowchart of the ASTD production pipeline.

transcribed sequences. These include alternative splicing databases such as ASAP II [5], ECGene [6], HOLLYWOOD [7], H-DBAS [8] and FAST DB [9], the FANTOM 3 database [10] that also features transcription initiation variants in the form of Cap-analysis gene expression (CAGE) tags, and the polyadenylation-specific PolyA_DB [3].

Here we report the development of ASTD, the Alternative Splicing and Transcript Diversity database, which aims at further integrating data from all three types of transcript variation together with extensive biological and expression information. ASTD is built upon and supersedes the splicing-oriented ASD database [11] and the ATD database that included 3′ end variations [12]. The Alternative Transcript Diversity Consortium that produced this new database involves laboratories working on various aspects of alternative transcript analysis, both computational and experimental. Our goal was to create an alternative transcript database that: (i) covers all three aspects of alternative transcription; (ii) includes three model vertebrate species and allows for expansion to include new species; (iii) offers a powerful interface for expression pattern-based queries; (iv) is fully integrated with other genomic data and genome browsing capabilities and (v) is extensible with respect to functional features and regulatory motifs.

Table 1
Release statistics for ASTD version 1.1

	Human	Mouse	Rat
Genes with an ASTD transcript	16,710	16,491	10,424
Genes with an ASTD transcription start site	13,265	11,161	1906
Genes with an ASTD polyA site	15,376	13,556	8842
Genes with an ASTD splice event	11,316	9474	2865
Genes with multiple TSSs	11,340	8150	4
Genes with splice events and multiple TSSs	10,145	8462	1
Genes showing splice events and multiple polyA sites	7607	4352	1275
Percentage of genes[a] undergoing alternative splicing	68%	57%	27%
Percentage of genes[a] undergoing alternative polyadenylation	92%	82%	85%
Percentage of genes undergoing multiple TSS positions	68%	49%	0.04%

[a] The number of protein coding genes with alternative forms as a percentage of the total number of protein coding genes in ASTD.

Results and discussion

Contents and comparisons with other transcript databases

The ASTD alternative transcript collection is built through three successive stages of transcript-to-genome mapping corresponding to

Table 2
Comparison of alternative splicing analyses with other databases

Database	Species	Genes alternatively spliced	Genes spliced	Exons
ASAP II[a]	Human	11,717 (53%)	22,220	129,981
ECGene[b]	Human	21,419 (45%)	47,943	–
Hollywood[c]	Human	–	–	151,199
ASTD	Human	14,101 (84%)	16,715	325,692
ASAP II	Mouse	8711 (53%)	16,404	105,260
ECGene	Mouse	19,361 (50%)	38,864	–
Hollywood	Mouse	–	–	90,885
ASTD	Mouse	13,028 (79%)	16,491	275,612
ASAP II	Rat	3378 (24%)	14,195	61,303
ECGene	Rat	11,005 (39%)	27,975	–
ASTD	Rat	6344 (61%)	10,424	122,593

[a] From ASAP II [5]. Genome assemblies: NCBI human Build 35 (UCSC version hg17), NCBI mouse Build 35 (UCSC version mm7), and RGSC rat Build 3.1 (UCSC version rn3).
[b] From the ECGene web site (http://genome.ewha.ac.kr/ECgene). Part A+B+C of database. Genome assemblies: NCBI human Build 36 (UCSC version hg18), NCBI mouse build 36 (UCSC version mm8), and RGSC rat Build 3.4 (UCSC version rn4).
[c] Exon number obtained from current statistics on http://hollywood.mit.edu/Logo/Fig_2.png. Genome assemblies: NCBI human Build 34 (UCSC version hg16) and NCBI mouse build 34 (UCSC version mm3).

splicing (AltSplice), polyadenylation (AltTrans and AltPAS) and transcriptional start site (AltTSS) variant prediction. AltTrans identifies polyadenylation sites corresponding to specific splice patterns while AltPAS identifies other potential polyA sites irrespective of underlying splice patterns. Each program maps a specific set of complementary DNA (cDNA) or expressed sequence tags (ESTs) to genome sequences, using protocols detailed in the Materials and methods section. The "Expression Profiles" pipeline associates each cDNA or sequence tag to anatomical/disease/development terms (expression states) and computes the expression profile of each alternative transcript based on numbers of cDNAs or tags supporting this transcript. Currently, expression profiles are computed for splice

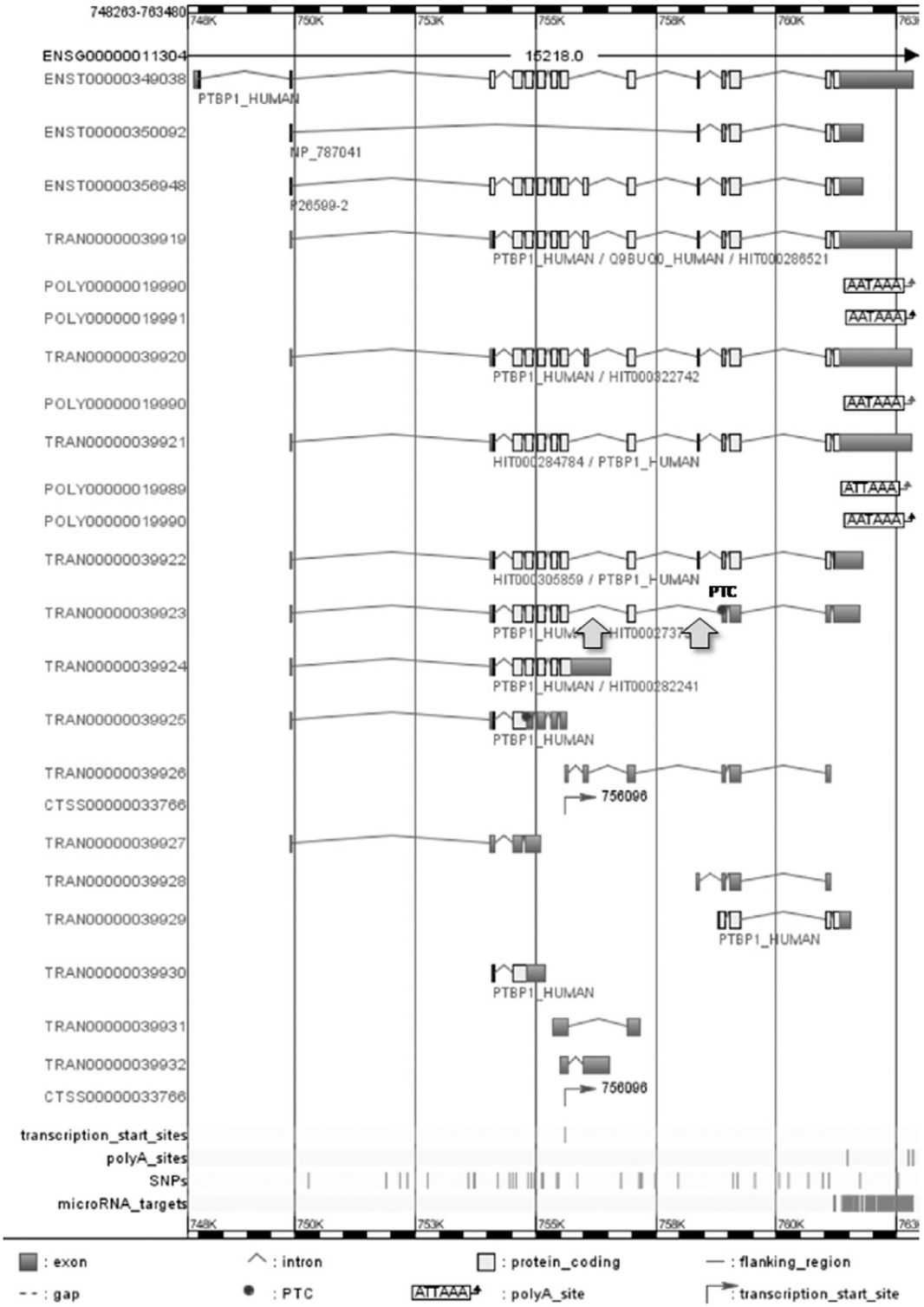


Fig. 2. Gene view for human polypyrimidine tract binding protein 1 gene PTBP1 (Ensembl ID ENSG00000011304.8). Each line presents a transcript isoform. The first three lines show transcript variants present in the Ensembl database and the next lines show ASTD variants. Alternative polyadenylation sites supported by cDNA data are presented with their respective signal sequence. Dark dots indicate premature termination codons present in internal exons that may lead to NMD. Transcript TRAN00000039923 has two skipped exons (arrows, described in literature as exons 9 and 11) and cause NMD.

variants only, but the procedure will eventually be extended to initiation and polyadenylation variants. The process is entirely automated, except for the derivation of expression states from novel cDNA/tag libraries that requires human curation. A flowchart presenting the general ASTD pipeline is shown in Fig. 1.

Table 1 presents the number of splicing, polyadenylation and transcriptional start site (TSS) variants per species in ASTD Version 1.1. The predicted fraction of genes with polyadenylation variants exceeds current estimates [3] mostly because ASTD also includes putative polyadenylation signals in the 10 kb downstream of region of reference gene models [13]. Splice and TSS variant frequencies are in line with current estimates, except for rat, which shows fewer events due to scarce EST/cDNA coverage in this species, especially for the 5′ region of genes. Table 2 compares ASTD splice variant statistics with corresponding values from three major AT databases. ASTD contains more alternative splicing events and/or mapped exons than ASAP II and Hollywood. This reflects multiple factors such as the use of more recent EST/genome database versions in ASTD and differences in the transcript mapping procedures. ECgene harbours significantly more genes and splice events than any other database due to their transcript construction procedure based on EST clustering. In Supplementary Table 1, we compare the numbers of alternatively spliced genes and proportions of different event types (cassette exons, alternative 5′ or 3′ sites, mutually exclusive exons) between the ASTD and ASAP II database. While ASTD has significantly higher splice event coverage in mouse and rat, event types are similarly distributed in the two databases: cassette exons are more frequent, followed by variations at donor/acceptor sites.

The ASTD database integrates Ensembl [14] features such as transcripts, exons and peptides, enabling comparison of ASTD and Ensembl predictions. Other transcript information and crosslinks include conserved splice junctions and splice events in human, mouse and rat; single nucleotide polymorphism (SNP) locations; transcripts with premature termination codons (PTCs) that may be subject to nonsense-mediated decay (NMD); microRNA targets; and peptide data, including for each variant the peptide sequence, domains and functional site signatures.

Display and query tools

ASTD offers multiple visualization levels, from complete chromosome map to the levels of genomic fragment, gene, transcript, splicing event and peptide. Users can easily navigate between different views, starting either from genomic coordinates or keyword search. The main visualization levels are the gene view showing all transcript isoforms, the transcript view displaying detailed exon information and the "event" view displaying independent alternative splicing events. Fig. 2 presents the gene view for the human polypyrimidine tract binding protein 1 gene (PTBP1), a splicing factor known to autoregulate its own splicing. A form lacking exons 9 and 11 is reported to undergo NMD and may contribute to the control of PTBP1 expression [15]. Indeed, this form appears to contain a premature termination codon as shown in Fig. 2 (TRANS00000039923).

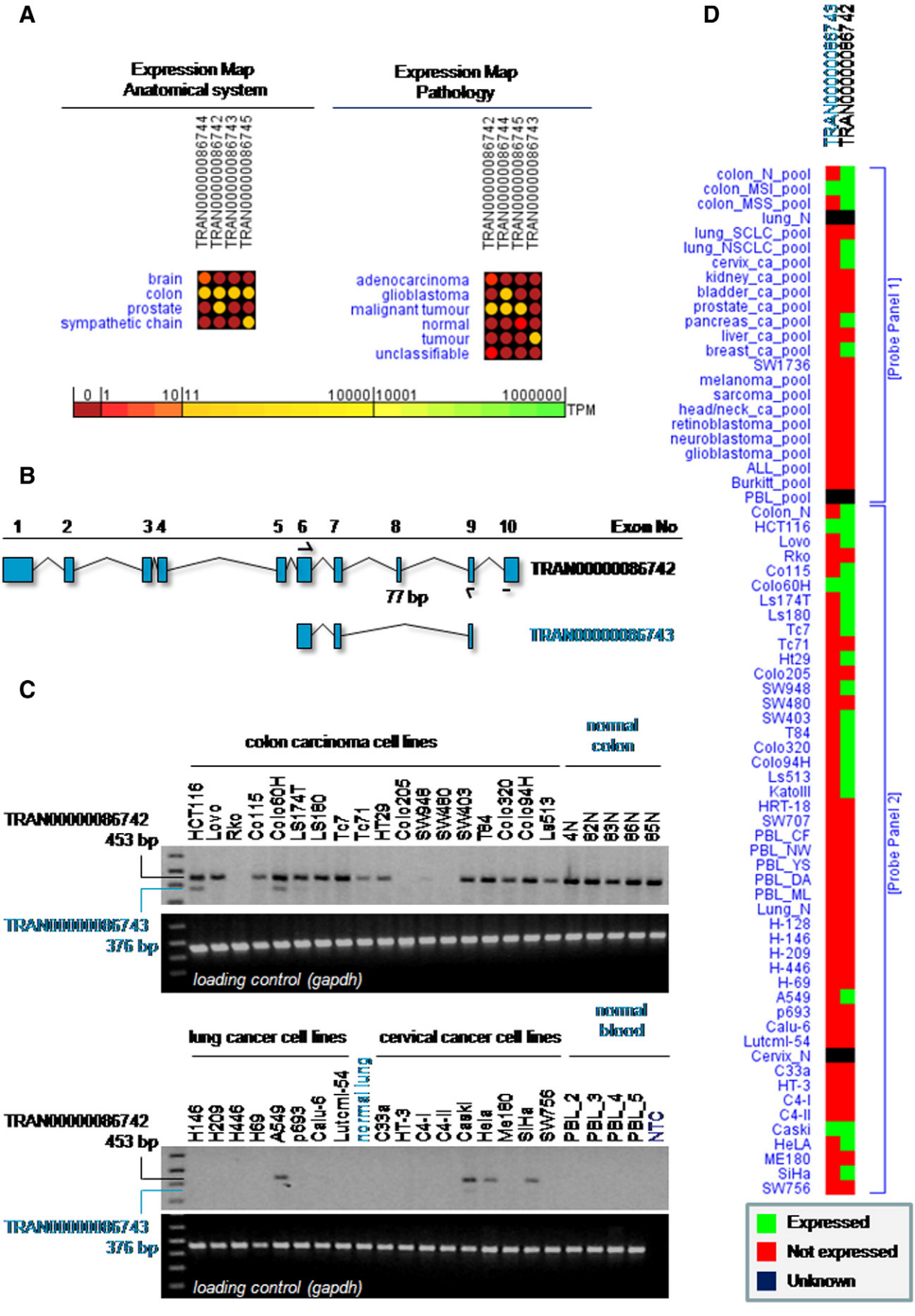
Expression state information is an important aspect of the ASTD database as it is used to analyze tissue-, developmental stage-, and disease-specific expression of alternative transcripts. Expression states in ASTD are derived from counts of cDNA/EST numbers in libraries (see Materials and methods). Results from such analyses are subject to well known limitations linked to cDNA library quality and construction methods [16,17]. However these analyses have repeatedly proven to be useful as indirect indications of expression biases [18,19,20,21,22]. The ASTD server provides various analysis tools that enable researchers to identify alternative transcripts of special biological interest, using two major query modes. First, from the main page, users can perform text search using ASTD identifiers, gene names, gene symbols, Ensembl IDs, UniProtKB entry names and accession numbers, UniProtKB/Swiss-Prot and UniProtKB/TrEMBL cross-references, EMBL evidence, tissues, pathologies, developmental stages, etc. Alternatively, an advanced search mode enables multi-criteria searches based on chromosome location, splicing events, GO (Gene Ontology) terms or gene expression patterns based on eVOC (a controlled vocabulary for unifying gene expression data [23]) annotation. In the latter "gene expression" mode, users can select two complex pools of tissues, developmental stage or pathologies, and obtain all alternative transcripts that are specifically expressed in one of the pools but not in the other. An overview of the expression states of all transcripts from a given gene can be displayed in the form of a digital expression map. This heatmap display enables a quick visual inspection of expression biases relative to anatomical systems, pathologies or developmental stages. Two expression maps (relative to anatomical system and pathology) for the human mucin 12, cell surface associated gene (MUC12) are shown in Fig. 3A. Users can also combine anatomy and pathology terms in a single expression map.

Experimental validation

Our project had a significant emphasis on the experimental validation of predicted transcripts using reverse-transcription (RT)-PCR. We completed the experimental validation of over 500 different polyadenylation [24] and splicing (results herein) events in human and mouse that were not previously described in the literature. For polyadenylation sites, our validation efforts focused on events conserved between human and mouse and producing alternative 3′ isoforms with size variations of 3 kb or more. Out of 86 such events, 84 were individually confirmed using a specially devised RT-PCR strategy [24]. For splice variants, we focused on the identification of cancer-specific events in human. From a list of 419 ASTD-predicted tissue-specific splice events for colorectal and/or lung cancer cells (Supplementary Table 6), we confirmed the existence of 370 independent events. With the objective of identifying new biomarkers for the detection of cancer cells, we further analyzed splice events occurring specifically in cancer cell lines/tissues. Among the validated splice events, 68 were detected in colorectal and/or lung cancer cell lines but not in normal colon mucosae and/or the analyzed normal lung. These cancer-specific splice events will represent a valuable source for further validation of colorectal and/or lung cancer biomarkers. Splice variant validation is presented in more detail in Supplementary Document 2.

Fig. 3 shows validation detail for the human MUC12 gene. The digital expression map obtained from the ASTD database (Fig. 3A) shows that while this gene is mainly expressed in the colon, transcript TRAN00000086743 is predicted to be colon tumor-specific. This transcript (Fig. 3B) lacks exon 8 relative to transcript TRAN00000086742 which should be present in colorectal normal and cancer cells. RT-PCR experiments confirmed these predictions. The transcript including exon 8 was mainly detectable in colon cell lines and normal colon tissues while the skipped exon transcript was present in colorectal cancer cell lines and

Fig. 3. Experimental validation of isoform expression for mucin 12, cell surface associated gene MUC12 (Ensembl ID ENSG00000169887.2). A: Digital expression profile of MUC12 transcripts (anatomical system and pathology). For each ASTD variant, TPM (transcript per million) expression values based on the numbers of supporting EST/cDNA from all relevant tissues are presented. Transcript TRAN00000086743 is predicted to be colon tumor-specific. B: Exon/intron structure of TRAN00000086742 and TRAN00000086743 and primer binding sites for RT-PCR validation. TRAN00000086743 has a skipped exon (exon 8; 77 bp). C: RT-PCR results in normal human tissues *vs.* human cancer cell lines of the colorectum, lung and cervix and healthy peripheral blood cells; PCR band 453 bp in length represents TRAN00000086742; TRAN00000086743 is confirmed by the 376 bp PCR fragment. A PCR system to amplify cDNAs of housekeeping glyceraldehyde-3-phosphate dehydrogenase gene GAPDH was used as loading control. NTC: non-tissue control (negative control without cDNA). D: RT-PCR Validation map of TRAN00000086742 and TRAN00000086743. The validation data is subdivided into a set of pooled human cell lines (probing panel I) on the right-hand side and a set of human colon, lung, cervical carcinoma cell lines (probing panel II) on the left-hand side.


A
Expression Map
Anatomical system
TRAN0000086744
TRAN0000086742
TRAN0000086743
TRAN0000086745
brain
colon
prostate
sympathetic chain
Expression Map
Pathology
TRAN0000086742
TRAN0000086744
TRAN0000086745
TRAN0000086743
adenocarcinoma
glioblastoma
malignant tumour
normal
tumour
unclassifiable
0 1 10 11 10000 10001 1000000
TPM
B
1 2 3 4 5 6 7 8 9 10 Exon No
TRAN0000086742
77 bp
TRAN0000086743
C
colon carcinoma cell lines
normal colon
HCT116 Lovo Rko Co115 Colo60H LS174T LS180 Tc7 Tc71 HT29 Colo205 SW948 SW480 SW403 T84 Colo320 Colo94H Ls513 4N B2N B3N B6N B5N
TRAN0000086742
453 bp
TRAN0000086743
376 bp
loading control (gapdh)
lung cancer cell lines
cervical cancer cell lines
normal blood
H146 H209 H446 H69 A549 p693 Calu-6 Lutcml-54 normal lung C33a HT-3 C4-I C4-II Caski Hela Me180 SiHa SW756 PBL_2 PBL_3 PBL_4 PBL_5 NTC
TRAN0000086742
453 bp
TRAN0000086743
376 bp
loading control (gapdh)
D
TRAN0000086743
TRAN0000086742
colon_N_pool
colon_MSI_pool
colon_MSS_pool
lung_N
lung_SCLC_pool
lung_NSCLC_pool
cervix_ca_pool
kidney_ca_pool
bladder_ca_pool
prostate_ca_pool
pancreas_ca_pool
liver_ca_pool
breast_ca_pool
SW1736
melanoma_pool
sarcoma_pool
head/neck_ca_pool
retinoblastoma_pool
neuroblastoma_pool
glioblastoma_pool
ALL_pool
Burkitt_pool
PBL_pool
[Probe Panel 1]
Colon_N
HCT116
Lovo
Rko
Co115
Colo60H
Ls174T
Ls180
Tc7
Tc71
Ht29
Colo205
SW948
SW480
SW403
T84
Colo320
Colo94H
Ls513
KatoIII
HRT-18
SW707
PBL_CF
PBL_NW
PBL_YS
PBL_DA
PBL_ML
Lung_N
H-128
H-146
H-209
H-446
H-69
A549
p693
Calu-6
Lutcml-54
Cervix_N
C33a
HT-3
C4-I
C4-II
Caski
HeLA
ME180
SiHa
SW756
[Probe Panel 2]
Expressed
Not expressed
Unknown

absent in all five tested normal colon samples (Fig. 3C). PCR validation data are incorporated in the ASTD database in the form of graphical overviews showing positive/negative PCR results for each tested splice variant and condition (Fig. 3D).

Availability and future directions

The ASTD database can be accessed at http://www.ebi.ac.uk/astd and is available for export as flat files containing all features under EMBL, GFF, GTF, BED, Excel and Fasta formats. Data can be obtained separately for each species. Update plans involve quarterly runs of the ASTD pipeline in order to represent newly generated transcribed sequences. Our objective is that within two years, the alternative transcript prediction, display and querying functionality of ASTD should be an integral part of the Ensembl database [14].

Materials and methods

Alternative transcripts are derived using a three-stage procedure. The genomic sequences used in ASTD 1.1 for *Homo sapiens*, *Mus musculus* and *Rattus norvegicus* are respectively based on NCBI human genome assembly build 36, NCBI mouse genome assembly build 36 and RGSC rat genome assembly build 3.4. The Ensembl gene sets referenced in ASTD 1.1 for *H. sapiens*, *M. musculus* and *R. norvegicus* come respectively from the Ensembl version 41, Ensembl version 41 and Ensembl version 42 gene builds.

Splicing variants

We generated splicing isoforms and alternative splice events using AltSplice, an automated computation pipeline for human, mouse and rat data [25]. Briefly, introns/exons, splicing isoforms and alternative splice events (cassette exons, mutually exclusive exons, etc.) are predicted through mapping of EST, cDNA and mRNA sequences from the standard, EST and high throughput cDNA (HTC) divisions of DDBJ/EMBL/GenBank [26] onto genomic sequences centred around Ensembl genes +/−10 kb. Mapped transcripts from EMBL release 88 include 8,125,884 human, 4,935,071 mouse and 824,394 rat sequences. Typically, less than 25% of mapped transcripts are retained as supports for splice variants in ASTD after ambiguous, incomplete and unspliced matches are discarded.

Polyadenylation variants

We identified polyadenylation variants and mapped them to individual splice variants as described in Le Texier et al. [12]. First, polyadenylation sites were identified at the 3′ end of each splice variant from the previous stage, requiring support by poly(A)/poly(T) terminated ESTs and the presence of a known poly(A) signal. Other poly(A) sites were predicted independently of splice variants based on a complete mapping of Genbank 3′ ESTs and full length cDNAs from H-InvDB [27] and FANTOM 3 [10] to genomic sequences. Criteria for inclusion of these splice variants include checks for internal priming sites (genomic poly(A) stretches), unmatched transcript ends and presence of a known polyadenylation signal, as described previously [28]. Poly(A) sites from both pipelines were then merged.

TSS variants

We identified TSSs for each transcript using oligo-capped full-length cDNA libraries. Transcript sequence sources are provided in Supplementary Table 2. Similarly to other studies ([29] and [30]), TSSs are defined as genomic positions matching the 5′ end of an oligo-capped cDNA and located either in the 5′ exon of a transcript or upstream. Here we used regions up to ten kilobases upstream of each splice variant. These regions were each aligned to the oligo-capped cDNA sequences using the NCBI-Blast program. The high-scoring segment pairs (HSP) with at least 95% identity were extracted and filtered according to the following additional criteria:

- Unambiguous match of cDNA to a single genomic region (the upstream-most HSP is considered, thereby defining the longest possible transcript);
- 3′ end of cDNA located downstream of the defined 5′ untranslated region (UTR);
- HSP covering >90% of cDNA.

When several TSS are present, we cluster TSSs into “promoter regions”. The distribution of inter-TSS distances is shown in Supplementary Fig. 3. Inter-TSS distances seldom exceed 500 bp in human and 300 bp in mouse. Based on this observation and on a previous analysis of 5′ end sequences in full-length cDNAs by Kimura et al. [29], we defined a promoter region as a TSS cluster with no gap over 500 nt. We did not cluster rat TSSs as most rat transcripts had a single TSS due to limited

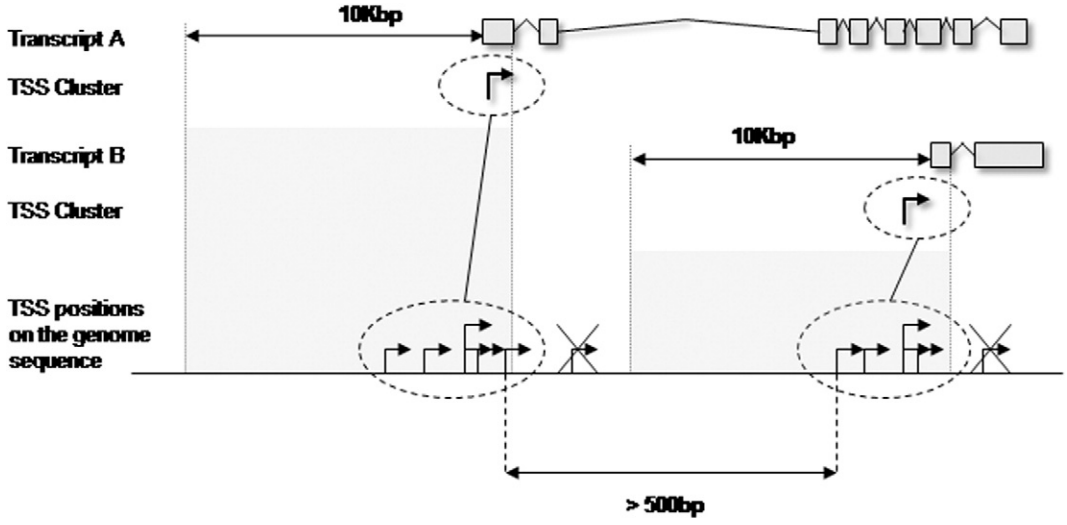


Fig. 4. TSS identification and clustering scheme. Two transcripts from a single gene are shown. Exons are represented by boxes and introns by lines. TSS clusters are represented with a thick arrow. TSSs are represented with a thin arrow. Groups of TSSs are identified in the 10 kbp UTR region (including first exon) of transcripts A and B based on the oligo-capped cDNA alignments. Crossed out TSS elements are discarded as they align outside of the first exon and 10 kb upstream region.

cDNA coverage. Fig. 4 gives an example of the identification and clustering of TSSs. The 3′-most TSSs of transcripts A and B were not considered as they aligned outside of the first exons and 10 kb upstream regions. TSS clustering then produced a single cluster for each transcript.

Procedures for alternative transcript derivation are detailed in Supplementary Document 1. Supplementary Table 3 summarizes the numbers of ESTs, cDNAs and mRNAs used to support the different transcript variants in each species.

Digital expression maps

Expression states were inferred from the clone library information associated with each mRNA or EST sequence that supported a given transcript. EST/cDNA sequences supporting several isoforms were assigned to each compatible isoform. Human and mouse expression states are based on the eVOC ontologies for anatomical systems, pathologies and development stages [23]. eVOC associates clone libraries and their corresponding sequences to controlled terms in an organized set of hierarchical vocabularies. Because there is no controlled vocabulary mapping for rat, as yet, we based our mapping on Medical Subject Headings (MeSH) developed by the US National Library of Medicine. We associated anatomy and disease vocabulary to rat cDNA libraries using a semi-automatic learning process. We grouped the developmental stages terms of rat and mapped them onto Witschi's embryonic development classification [31] instead of the Theiler mouse-specific stages. We dealt with any ambiguity in library terms using a custom-made dictionary where enough information was available. General anatomical terms such as “head” or whole tissue terms like “brain” constituted of several cell types were associated to the equivalent or most specific term in the hierarchy. In all species, ESTs or cDNAs from mixed or pooled tissue libraries were discarded.

Supplementary Table 4 provides numbers of analyzed libraries and extracted ontology terms for the three species studied. Tissue representation in EST/cDNA libraries from each organism is presented in Supplementary Figs. 1 and 2. As expected, some tissues are covered very differently in different species. For instance, lung is among the top ten tissues in human and rat while it is seldom represented in mouse, in terms of number of libraries (Supplementary Fig. 1) as well as of number of ESTs (Supplementary Fig. 2).

We derived expression data for three types of expression states described in cDNA libraries, namely anatomy, pathology and development. We used normalized transcript digital expression values, introduced by the NCBI to mine UniGene for tissue/disease specificity [32]. These values are expressed in TPM (transcript per million). The TPM values of transcripts are calculated as follows: for any transcript, divide the number of EST evidences found in a particular expression state (*e.g.* “lung”) by the total number of ESTs from the cDNA libraries related to this expression state. The resulting value is then normalized to one million.

We measured fold changes in the expression of a transcript in a given expression state by dividing TPM in this state by the average TPM expression value of this transcript in all expression states. The digital differential expression significance was then measured using a *t*-test with a *p*-value cut-off of 0.05, applying a Benjamini–Hochberg correction for a false discovery rate of less than 5%. This produced a final set of 38%, 22% and 19% significant differentially expressed isoforms for human, mouse and rat, respectively, between two conditions (for example, normal *vs.* cancer). For each transcript, an “expression analysis” page (Supplementary Fig. 4) displays TPM, fold change between states and *p*-value of differential expression for each type of expression state (anatomy, development and pathology).

Derivation of other data

Conserved alternative splicing

ASTD identifies pairs of orthologous transcripts in any combination of human, mouse or rat displaying conserved splice junctions and conserved splice events, as defined by Thanaraj et al. [33]. First, we identify conserved splice junctions that are conserved between genes in the human and mouse AltSplice data sets by examining mouse AltSplice transcript sequences (that are already mapped to a mouse gene) for homology to exon sequence constructs around human AltSplice splice junctions. An exon sequence construct is built by concatenating two components: the 5′ component being the 70 nt sequence from the 3′ end of the 5′ exon and the 3′ component being the 70 nt sequence from the 5′ end of the 3′ exon. A mouse gene whose transcriptome data shows at least three splice junctions conserved with those from the human gene is considered as orthologous to the human gene in question. Similarly, mouse transcript patterns are considered as orthologous to a human transcript pattern if at least three splice junctions are conserved between the two.

SNPs

We identified SNP positions based on dbSNP [34] as previously reported in [12]. We aligned SNPs from the dbSNP databank to each ASTD transcript to identify exonic SNPs and their allele usage by examining each corresponding nucleotide of the EST/mRNA that supported the transcript sequence. We derived allele-usage frequencies at each SNP position and deduced SNPs with significant differences in allele usage among the transcripts. Such SNPs may therefore mediate the regulation of alternative splicing. Users can visualize SNPs associated to each transcript variant.

Peptides

For each variant transcript we derived the corresponding peptide sequences, domain and functional site signatures obtained from InterProScan [35] as well as synonymous and non-synonymous SNPs. The web interface presents the longest CDS of each splice variant by default. A link to the other possible translations is provided (note that the other CDS are not necessarily in frame with the longest CDS for the entire gene).

Premature termination codons (PTC)

In the coding sequence of ASTD transcripts, we annotate as PTC nonsense codons located more than 50 nucleotides upstream of any exon–exon junction [36].

MicroRNA targets

Target sites were obtained from the ElMMo miRNA target prediction server (inference method developed by Gaidatzis et al. [37]). Because microRNA targets were originally predicted on mRNA sequences from NCBI RefSeq, we mapped the target positions on the mRNA sequences back to the gene sequences in ASTD.

Experimental validation

We confirmed alternative polyadenylation events using an *ad-hoc* PCR strategy. Briefly, we extracted mRNAs from 25 mouse tissues and cell lines and we designed three specific RT-PCR probe sets for the amplification of: (i) independent 3′ ends; (ii) tandem 3′ ends and (iii) very long 3′ UTRs (over 4.5 kb). Detailed protocols are given in [24]. For alternative splicing events validation, we obtained mRNAs from 19 human tissues and 95 cell lines. Existence and expression profile of selected events for validation were analyzed with specific RT-PCR using cDNAs from two different tissue/cell line panels. The first panel consisted of pooled cell lines representing 19 different tissue origins. When splice events were detected, a second probing panel was analyzed. PCR product integrity and fragment sizes were verified on agarose gels. A detailed protocol is provided in Supplementary Document 2.

Acknowledgments

This work was funded by the European Commission FP6 Programme, contract number LSHG-CT-2003-503329.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2008.11.003.

References

[1] F. Denoeud, et al., Prominent use of distal 5′ transcription start sites and discovery of a large number of additional exons in ENCODE regions, Genome Res. 17 (2007) 746–759.
[2] M.L. Tress, et al., The implications of alternative splicing in the ENCODE protein complement, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 5495–5500.
[3] J.Y. Lee, I. Yeh, J.Y. Park, B. Tian, PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes, Nucleic Acids Res. 35 (2007) D165–D168.
[4] A. Srebrow, A.R. Kornblihtt, The connection between splicing and cancer, J. Cell. Sci. 119 (2006) 2635–2641.
[5] N. Kim, A.V. Alekseyenko, M. Roy, C. Lee, The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species, Nucleic Acids Res. 35 (2007) D93–D98.
[6] Y. Lee, et al., ECgene: an alternative splicing database update, Nucleic Acids Res. 35 (2007) D99–D103.
[7] D. Holste, G. Huo, V. Tung, C.B. Burge, HOLLYWOOD: a comparative relational database of alternative splicing, Nucleic Acids Res. 34 (2006) D56–D62.
[8] J. Takeda, et al., H-DBAS: alternative splicing database of completely sequenced and manually annotated full-length cDNAs based on H-Invitational, Nucleic Acids Res. 35 (2007) D104–D109.
[9] P. de la Grange, M. Dutertre, N. Martin, D. Auboeuf, FAST DB: a website resource for the study of the expression regulation of human gene products, Nucleic Acids Res. 33 (2005) 4276–4284.
[10] P. Carninci, et al., The transcriptional landscape of the mammalian genome, Science 309 (2005) 1559–1563.
[11] S. Stamm, et al., ASD: a bioinformatics resource on alternative splicing, Nucleic Acids Res. 34 (2006) D46–D55.
[12] V. Le Texier, et al., AltTrans: transcript pattern variants annotated for both alternative splicing and alternative polyadenylation, BMC Bioinformatics 7 (2006) 169.
[13] F. Lopez, S. Granjeaud, T. Ara, B. Ghattas, D. Gautheret, The disparate nature of "intergenic" polyadenylation sites, RNA 12 (2006) 1794–1801.
[14] P. Flicek, et al., Ensembl 2008, Nucleic Acids Res. 36 (2008) D707–D714.
[15] M.C. Wollerton, C. Gooding, E.J. Wagner, M.A. Garcia-Blanco, C.W. Smith, Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay, Mol. Cell. 13 (2004) 91–100.
[16] R. Sorek, H.M. Safer, A novel algorithm for computational identification of contaminated EST libraries, Nucleic Acids Res. 31 (2003) 1067–1074.
[17] S. Gupta, D. Zink, B. Korn, M. Vingron, S.A. Haas, Strengths and weaknesses of EST-based prediction of tissue-specific alternative splicing, BMC Genomics 5 (2004) 72.
[18] Z. Wang, et al., Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer, Cancer Res. 63 (2003) 655–657.
[19] Q. Xu, C. Lee, Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences, Nucleic Acids Res. 31 (2003) 5635–5643.
[20] L. Hui, et al., Identification of alternatively spliced mRNA variants related to cancers by genome-wide ESTs alignment, Oncogene 23 (2004) 3013–3023.
[21] H. Zhang, J.Y. Lee, B. Tian, Biased alternative polyadenylation in human tissues, Genome Biol. 6 (2005) R100.
[22] W. Ritchie, S. Granjeaud, D. Puthier, D. Gautheret, Entropy measures quantify global splicing disorders in cancer, PLoS Comput. Biol. 4 (2008) 3.
[23] J. Kelso, et al., eVOC: a controlled vocabulary for unifying gene expression data, Genome Res. 13 (2003) 1222–1230.
[24] V. Moucadel, F. Lopez, T. Ara, P. Benech, D. Gautheret, Beyond the 3′ end: experimental validation of extended transcript isoforms, Nucleic Acids Res. 35 (2007) 1947–1957.
[25] T.A. Thanaraj, et al., ASD: the Alternative Splicing Database, Nucleic Acids Res. 32 (2004) D64–D69.
[26] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, D.L. Wheeler, GenBank, Nucleic Acids Res. 36 (2008) D25–D30.
[27] T. Imanishi, et al., Integrative annotation of 21,037 human genes validated by full-length cDNA clones, PLoS Biol. 2 (2004) e162.
[28] E. Beaudoing, D. Gautheret, Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data, Genome Res. 9 (2001) 1520–1526.
[29] K. Kimura, et al., Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes, Genome Res. 16 (2006) 55–65.
[30] R. Yamashita, et al., DBTSS: DataBase of human transcription start sites, progress report 2006, Nucleic Acids Res. 34 (2006) D86–D89.
[31] E. Witschi, Development; rat, in: P.L. Altman, D.S. Dittmer (Eds.), Growth Including Reproduction and Morphological Development. Biological Handbooks of the Federation of American Societies for Experimental Biology, Washington, DC., 1962, pp. 304–314.
[32] The UniGene Digital Differential Display [http://www.ncbi.nlm.nih.gov/UniGene/ddd.cgi].
[33] T.A. Thanaraj, F. Clark, J. Muilu, Conservation of human alternative splice events in mouse, Nucleic Acids Res. 31 (2003) 2544–2552.
[34] E.M. Smigielski, K. Sirotkin, M. Ward, S.T. Sherry, dbSNP: a database of single nucleotide polymorphisms, Nucleic Acids Res. 28 (2000) 352–355.
[35] N.J. Mulder, et al., New developments in the InterPro database, Nucleic Acids Res. 35 (2007) D224–D228.
[36] L.E. Maquat, Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics, Nat. Rev. Mol. Cell. Biol. 5 (2004) 89–99.
[37] D. Gaidatzis, E. van Nimwegen, J. Hausser, M. Zavolan, Inference of miRNA targets using evolutionary conservation and pathway analysis, BMC Bioinformatics 8 (2007) 69.