**Original article**

**DBATE: database of alternative transcripts expression**

Valerio Bianchi¹, Alessio Colantoni¹, Alberto Calderone², Gabriele Ausiello¹, Fabrizio Ferre¹,* and Manuela Helmer-Citterich¹

¹Centre for Molecular Bioinformatics, Department of Biology, University of Rome Tor Vergata, Via della Ricerca Scientifica s.n.c., Rome 00133, Italy and ²Department of Biology, University of Rome Tor Vergata, Via della Ricerca Scientifica s.n.c., Rome 00133, Italy

Present address: Valerio Bianchi, Center for Genomic Science of IIT@SEMM, Istituto Italiano di Tecnologia, via Adamello 16, Milan 20139, Italy.

*Corresponding author: Tel: +39 06 7259 4320; Fax: +39 06 2023 500; Email: fabrizio.ferre@uniroma2.it

Submitted 21 November 2012; Revised 7 June 2013; Accepted 9 June 2013

Citation details: Bianchi, V., Colantoni, A., Calderone, A., et al. DBATE: database of alternative transcripts expression. (2013) Vol. 2013: article ID bat050; doi:10.1093/database/bat050.

The use of high-throughput RNA sequencing technology (RNA-seq) allows whole transcriptome analysis, providing an unbiased and unabridged view of alternative transcript expression. Coupling splicing variant-specific expression with its functional inference is still an open and difficult issue for which we created the DataBase of Alternative Transcripts Expression (DBATE), a web-based repository storing expression values and functional annotation of alternative splicing variants. We processed 13 large RNA-seq panels from human healthy tissues and in disease conditions, reporting expression levels and functional annotations gathered and integrated from different sources for each splicing variant, using a variant-specific annotation transfer pipeline. The possibility to perform complex queries by cross-referencing different functional annotations permits the retrieval of desired subsets of splicing variant expression values that can be visualized in several ways, from simple to more informative. DBATE is intended as a novel tool to help appreciate how, and possibly why, the transcriptome expression is shaped.

Database URL: http://bioinformatica.uniroma2.it/DBATE/.

**Introduction**

Alternative splicing (AS) permits the synthesis of multiple transcript variants from a single gene, thus increasing the diversity of RNAs and proteins encoded by a genome (1, 2). The number of known splicing variants in the human transcriptome stored in Ensembl is growing at a dramatic pace. Through the use of recent high-throughput RNA sequencing technologies (RNA-seq), it has been demonstrated that ~95% of multi-exon genes undergo AS in panels of human tissues (3), shaping the expressed transcriptome in various ways (4) and generating an exceedingly complex repertoire of mRNAs (5). Splicing variant expression deconvolution algorithms, such as Cufflinks (6), IsoEM (7), Scripture (8), RSEM (9) and SpliceSeq (10), allow the reliable (as validated in many instances by RT-PCR) quantitative estimation of the transcription of individual splicing variants of a gene from RNA-seq data. Yet, the functional interpretation of such expression data, or the change of splicing variant-specific expression patterns in different tissues or conditions, is still overly difficult. In the simplest cases, splicing promotes the inclusion or removal of specific exons corresponding to whole-protein domains to which a specific function can be assigned, but often splicing patterns are much more complex and the effect of splicing on the protein product function(s) is much more elusive. A number of indications suggest that in a considerable fraction of cases, splicing can radically change the protein product function and/or fold (11–14), and a non-negligible amount of splicing variants shows structural inconsistencies (e.g. low degrees of residue packing in the protein core or large fractions of hydrophobic residues exposed to the solvent), and lack of known
functional regions. As a consequence, despite the large amount of data available about AS variants and protein functional annotations, there are no resources dedicated to the integrated retrieval of such information.

A number of databases offer storage or download of next-generation sequencing data (15, 16), but the splicing variant-level expression analysis is still unfriendly for the biomedical researchers, given the exceedingly large amount of data to be processed, the computational power required and the nature of the analysis algorithms that are usually intended for the computational biologist. A small number of recent databases storing RNA-seq expression data only provide gene-level expression values (e.g. 17). Splicing variant-level annotations are starting to be available in databases, such as Uniprot, but they can be of difficult interpretation without a reference context, and are still largely incomplete. SpliceSeq (10) provides a user-friendly interactive graphic environment, integrated with isoform-specific functional annotations from Uniprot, but splicing variant-level expression estimation must be run by the user. Many databases exist that collect AS variants (18–21), but they do not tackle variant functional annotation. There are no general tools that can be used to infer whether a given variant is actually translated, and its eventual protein product stable and containing functional regions and residues. Various resources have been developed for the analysis of specific effects of AS, for example ProSAS (22) for the analysis of the changes introduced by AS on protein structures, or AS-ALPS and AS-EAST (23, 24) for the analysis of the effect of AS on protein–protein interfaces and other structure-based functional assignments. A web server, MAISTAS (25), provides a framework to test the structural consistency of a splicing variant, but has a limited range of application because it requires a high sequence identity between the variant under analysis and a template with known 3D structure. As a consequence, the integration of transcript-level RNA-seq expression and their functional characterization must be currently approached by combining different tools and cross-referencing heterogeneous databases and data types.

We aim at filling this void with the DataBase of Alternative Transcripts Expression (DBATE). We processed 13 large public RNA-seq panels from human healthy tissues and in disease conditions. For each splicing variant in each sample, we report the estimated transcript expression and its functional annotations, extracted and integrated from different sources: Ensembl (26), Pfam (27), Uniprot/Swiss-Prot (28), GO (29) and mentha (Calderone & Cesareni, in press; http://mentha.uniroma2.it/). The user can access splicing variant expression levels of the genes or transcripts of interest, compare them among different samples and perform more complex queries by cross-referencing the available annotations. The interface is designed to facilitate the data retrieval, available in five different formats: HTML tables, Excel spreadsheets, plain text tab-separated files, barplots and heatmaps.

DBATE content

DBATE provides the expression level for each human transcript annotated in Ensembl (release 67) estimated in 13 different panels of human tissues/cell lines available in the Gene Expression Omnibus (GEO) (30), enriched with functional annotation. These panels have been chosen to cover the largest number of samples from human healthy tissues, organs or cell lines; for seven of them, normal and tumoral condition is provided. Each sample in each panel was processed independently, and we provide tools for the comparison of any given set of samples that the user can select as desired. The list of available data sets is provided in Table 1 reporting, for each data set, its GEO identifier, the literature reference (when available).

RNA-seq analysis

All the data sets have been checked for read quality using FASTQC (v0.10.1 at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). We trimmed read ends if base quality scores were <20. At the end of this process, if the total length of the read is <10 bases, we discarded the read.

We used the Tuxedo Suite, comprising Bowtie (v0.12.7) (42), TopHat (v1.4.1) (43) and Cufflinks (v1.3.0) (6), to align the sequence reads produced in each experiment to the reference human genome hg19 seeking only unique genome matches with up to two mismatches (Bowtie), to identify splicing junctions (TopHat), and to evaluate the normalized expression of individual splicing variants (Cufflinks) reported in Fragments Per Kilobase of transcript per Million mapped reads (FPKM). Cufflinks is the de facto standard algorithm for the isoform deconvolution problem, and it is widely used (44). More recent algorithms are available (45–48), but in absence of appropriate benchmarks for this kind of algorithms we deemed more appropriate to select the most popular tool. We actually tested two additional splicing variant expression estimation algorithms, IsoEM (7) and Scripture (8), on the Wang data set (see above), finding a high correlation with the Cufflinks expression estimates (Pearson correlation coefficient >0.8 for both algorithms).

Functional annotation

Various sources of functional information have been integrated in DBATE, gathering functional annotations from the Ensembl, Uniprot/Swiss-Prot, Pfam, Gene Ontology and mentha databases. With the exception of protein interaction information from mentha (described later in the article), all annotations are mapped to individual splicing
### Table 1. Data sets included in DBATE

| GEO GSE identifier | Samples                                                                 | Number of reads ($>10^6$) | Read length (bp) | Description                                                                                     | Reference |
|--------------------|------------------------------------------------------------------------|-----------------------------|------------------|------------------------------------------------------------------------------------------------|----------|
| GSE12946$^a$      | Adipose, brain, breast, colon, heart, liver, lymph node, skeletal muscle, testes, BT474, HME, MB435, MCF-7, T47D | 224                         | 32$^a$           | The Wang data set, from which we selected 14 samples, 9 in normal condition and 5 in tumoral condition | 31       |
| GSE17274$^b$      | Three female (HSF1, HSF2, HSF3) and three male liver samples (HSM1, HSM2, HSM3) | 72                          | 35$^a$           | Sex-specific gene expression in liver in three males and three females                           | 32       |
| GSE29119$^b$      | Breast cancer (HCC1954) and normal breast cells (HMEC)                 | 97                          | 36$^a$           | Gene expression analysis of breast cancer                                                        | 33       |
| GSE29155$^a$      | Prostate epithelial (PrEC) and prostate adenocarcinoma (LNCaP) cell lines | 9                           | 36$^a$,40$^f$    | Transcription profiling of human prostate epithelial and adenocarcinoma cell lines              | 34       |
| GSE29580$^b$      | Normal and tumor samples from two colorectal cancer patients           | 40                          | 36$^a$           | Whole transcriptome sequencing of colorectal cancer                                              | NA       |
| GSE29968$^b$      | Matched esophageal squamous cells from three carcinoma patients         | 118                         | 38$^a$           | Transcriptome analysis of human esophageal squamous cell carcinoma in three pairs of matched patient-derived tumor samples and their adjacent non-tumorous tissues | 35       |
| GSE30611$^c$      | Adipose, adrenal, brain, breast, colon, heart, kidney, liver, lung, lymph node, ovary, prostate, skeletal muscle, testes, thyroid, white blood cells | 2000                        | 50$^f$           | The Illumina BodyMap 2.0 Project, comprising transcription profiling of individual and mixtures of 16 human tissues | NA       |
| GSE30772$^c$      | Mitochondrion, mitoplast                                            | 45                          | 35$^a$           | Examination of the mitochondrial transcriptome                                                   | 36       |
| GSE32689$^b$      | Pooled oocytes, pooled sister polar bodies, single oocyte, single sister polar body | 120                         | 42$^e$           | Transcriptome of the human polar body, providing four conditions: pooled oocytes and their sister polar bodies and a single oocyte and its sister polar body | 37       |
| GSE33328$^d$      | Peripheral brain tissue, tumor brain tissue                         | 49                          | 75$^a$           | Transcriptomic profiling of a glioblastoma multiforme patient with control peripheral brain tissue | 38       |
| GSE37769$^c$      | THP1 cells                                                            | 287                         | 100$^a$          | Expression analysis of the THP1 (human monocytic leukemia) cell line                            | 39       |
| GSE38685$^b$      | Prostate epithelial (PrEC) and prostate adenocarcinoma (LNCaP) cell lines | 35                          | 75$^a$           | Transcription profiling of human prostate epithelial and adenocarcinoma cell lines              | 40       |
| GSE43925$^b$      | THP1 high glucose, THP1 normal glucose                               | 60                          | 42$^e$           | Expression analysis of human THP-1 monocytes in normal conditions and treated with high glucose | 41       |

$^a$Platform: Genome Analyzer  
$^b$Platform: Genome Analyzer IIx  
$^c$Platform: HiSeq 2000  
$^d$Platform: Genome Analyzer II  
$^e$Single end reads  
$^f$Paired end reads

The current DBATE release includes 13 data sets retrieved from the Gene Expression Omnibus (GEO). The Table report for each data set its GEO GSE identifier, the samples it contains, the total number of reads (expressed in million reads), the read length, a brief description of the data set content, and the literature reference when available (NA indicates that the data were deposited in GEO but the study is still unpublished). Superscripts indicate the sequencing technology employed used (either GA, GAII, GAIIx, or HiSeq 2000), and whether the reads were sequenced as single or paired ends.
variants. Each feature can be used to filter the input query, selecting only those genes having at least one splicing variant carrying the desired annotation or set of annotations. It should be noted that, although the annotation transfer pipeline is intended for protein-coding transcripts, DBATE also stores expression levels for non-coding RNAs, which are assuming a central role in many cellular processes (49, 50).

We extracted from the Ensembl database (release 67) the definition (i.e. the exon composition) of each transcript linked to its gene (ENSG), transcript (ENST), associated protein (ENSP) identifiers and its genomic coordinates, defined by chromosome, absolute start–end on the human genome and strand. In total, DBATE stores 22 403 protein-coding Ensembl genes and their associated 100 357 transcripts, for which 94 737 (94.4%) encode for a different protein sequence (the remaining variants differ only in the untranslated regions). Additionally, DBATE stores 96 136 non-coding transcripts, some of which (52%) are non-coding isoforms of protein-coding genes, whereas the remaining 48% are products of genes for which no transcript is translated that might encode for functional RNAs, i.e. RNAs that are not translated because they have no ORF, but exert their function as RNA molecules (51, 52). A rapidly increasing interest in these non-coding RNA (ncRNAs) genes motivated their inclusion in DBATE, even if obviously in these cases we cannot apply our protein annotation transfer pipeline because these genes do not encode for proteins.

We collected all the human entries of the Uniprot/Swiss-Prot database resulting in 20 231 different entries. For each entry, we gathered the Uniprot ID, Uniprot Primary Accession Number, sequence of the main splicing variant, protein existence evidence, features and post-translational modifications. All the information stored in Uniprot is linked to its gene (ENSG), transcript (ENST), associated protein (ENSP) identifiers and its genomic coordinates, defined by chromosome, absolute start–end on the human genome and strand. In total, DBATE stores 22 403 protein-coding Ensembl genes and their associated 100 357 transcripts, for which 94 737 (94.4%) encode for a different protein sequence (the remaining variants differ only in the untranslated regions). Additionally, DBATE stores 96 136 non-coding transcripts, some of which (52%) are non-coding isoforms of protein-coding genes, whereas the remaining 48% are products of genes for which no transcript is translated that might encode for functional RNAs, i.e. RNAs that are not translated because they have no ORF, but exert their function as RNA molecules (51, 52). A rapidly increasing interest in these non-coding RNA (ncRNAs) genes motivated their inclusion in DBATE, even if obviously in these cases we cannot apply our protein annotation transfer pipeline because these genes do not encode for proteins.

We collected all the human entries of the Uniprot/Swiss-Prot database resulting in 20 231 different entries. For each entry, we gathered the Uniprot ID, Uniprot Primary Accession Number, sequence of the main splicing variant, protein existence evidence, features and post-translational modifications. All the information stored in Uniprot is linked to its gene (ENSG), transcript (ENST), associated protein (ENSP) identifiers and its genomic coordinates, defined by chromosome, absolute start–end on the human genome and strand. In total, DBATE stores 22 403 protein-coding Ensembl genes and their associated 100 357 transcripts, for which 94 737 (94.4%) encode for a different protein sequence (the remaining variants differ only in the untranslated regions). Additionally, DBATE stores 96 136 non-coding transcripts, some of which (52%) are non-coding isoforms of protein-coding genes, whereas the remaining 48% are products of genes for which no transcript is translated that might encode for functional RNAs, i.e. RNAs that are not translated because they have no ORF, but exert their function as RNA molecules (51, 52). A rapidly increasing interest in these non-coding RNA (ncRNAs) genes motivated their inclusion in DBATE, even if obviously in these cases we cannot apply our protein annotation transfer pipeline because these genes do not encode for proteins.

We collected all the human entries of the Uniprot/Swiss-Prot database resulting in 20 231 different entries. For each entry, we gathered the Uniprot ID, Uniprot Primary Accession Number, sequence of the main splicing variant, protein existence evidence, features and post-translational modifications. All the information stored in Uniprot is linked to its gene (ENSG), transcript (ENST), associated protein (ENSP) identifiers and its genomic coordinates, defined by chromosome, absolute start–end on the human genome and strand. In total, DBATE stores 22 403 protein-coding Ensembl genes and their associated 100 357 transcripts, for which 94 737 (94.4%) encode for a different protein sequence (the remaining variants differ only in the untranslated regions). Additionally, DBATE stores 96 136 non-coding transcripts, some of which (52%) are non-coding isoforms of protein-coding genes, whereas the remaining 48% are products of genes for which no transcript is translated that might encode for functional RNAs, i.e. RNAs that are not translated because they have no ORF, but exert their function as RNA molecules (51, 52). A rapidly increasing interest in these non-coding RNA (ncRNAs) genes motivated their inclusion in DBATE, even if obviously in these cases we cannot apply our protein annotation transfer pipeline because these genes do not encode for proteins.
one term in the ‘molecular function’ domain; 51 634 trans-
scripts are associated with at least one term in the ‘cellular
component’ domain. Ensembl and Biomart recently started
to associate GO terms with individual splicing variants.
Although such annotations are largely incomplete, yet
they provide in many cases variant-specific information.
GO term annotations will be frequently updated over time.

The mentha interactome database collected protein–
protein interactions (PPI) retrieved from five different PPI
databases—IntAct (53), MINT (54), DIP (55), BioGRID (56)
and MatrixDB (57)—with the aim of eliminating redun-
dancy between these different sources. The main moti-
vation of the mentha database is that currently available
curated databases of protein interactions offer only a lim-
ited view of the interactome which can be expanded and
made more consistent by their integration. We have com-
bined this database with DBATE through a Java Applet to
browse PPIs using the unique Uniprot ID associated to each
transcript.

Querying DBATE

The DBATE database archives a wide variety of functional
annotations. The user can access the splicing variant expres-
sion level of genes or transcripts of interest, or perform
more complex queries using an advanced form through
the use of cross-referenced annotations.

The DBATE user interface is intended to provide the
choice of increasingly complex queries, in an intuitive fash-
on. In the simplest query, the user inputs a transcript ID, and
retrieves pre-computed expression levels in FPKM units for
all tissue samples in BodyMap (the default sample group), a
simple TAB delimited plain-text file, a Microsoft Excel table
file and an HTML table. All functional annotations of the
input transcript are reported and organized in panels.
When the user inputs a gene ID or a gene name, all its spli-
cing variants are returned. Moreover, the database can be
accessed via one or more GO terms or Pfam domain IDs;
mixed queries are also allowed. Individual variants can be
manually selected from the list of variants matching the
query. In addition to the different tabular outputs and an-
notation panels, if the number of variants is higher than one
(and lower than 100), DBATE also offers a heatmap group-
ing similar expression patterns across all selected samples.

Using the advanced options input form, the user can create
more complex queries. First, specific samples from
all RNA-seq data sets can be selected. Then, functional an-
notations can also be chosen to filter the input transcripts
for only those matching the selected features. As a case
study, we report the analysis of proteins containing the K
Homology (KH) domain (Pfam ID: PF00013), a domain able
to bind RNA promoting its degradation and that is involved
in splicing regulation (58, 59). The KH domain has been
found in some cases associated to repeated protein
sequences such as the ankyrin repeat (60, 61), and its phos-
phorylation can modulate its binding ability (62). DBATE
can be easily queried to retrieve all splicing variants that
encode for protein products that contain the KH domain;
them the query can be refined to retain only splicing vari-
ants encoding for phosphorylation sites and containing
repeat units in their protein sequences, and for retrieving
expression values in the desired pool of samples. Such a
composite query that cross-links different types of informa-
tion retrieves 10 transcripts belonging to 4 different genes:
ANKRD17, KHSRP, HNRNPK and ANKHD1. Their expression
patterns are reported in the heatmap in Figure 1, showing
that splicing variants for these proteins can have different
and tissue-specific expression. Interestingly, not all the
ANKHD1 variants contain the KH domain. An ANKHD1 spli-
cing variant lacking the KH domain, which has important
roles in apoptosis, is reported in the literature (63).

Retrieving the full list of which ANKHD1 variants contain
the KH domain, and their expression patterns, is not a triv-
ial task, but can be immediately obtained from DBATE by
simply querying the ANKHD1 gene. DBATE reports that 10
out of 19 ANKHD1 splicing variants lack the KH domain,
and their expression patterns can help in elucidating their
cellular roles.

For each query, protein interaction data from the
mentha browser can be visualized through a Java applet
integrated in the results page. For each splicing variant
derived from the query, a list of unique Uniprot primary
accession numbers is used to interrogate mentha. For
each query protein, all physically interacting proteins
retrieved in mentha are reported as connected by an
edge to the query proteins, and the expression FPKM in a
selected tissue of each transcript is reported, for both the
query proteins and their binding partners. Each network
node is color-coded by the expression level of the dominant
isoform, whereas clicking the single node reports all the
different splicing variants with their expression values.
The mentha browser additionally allows expansion and
pruning of the network. The interaction network for the
4 KH domain-containing genes selected in the previous
paragraph is reported in Figure 2, where they display
close connectivity mediated by common binding partners.

Finally, expression data from all data sets and transcripts
can be downloaded as static tab-separated text files.

Data organization and web interface

DBATE has been implemented within the MySQL database
management system version 5.1.17 on a Linux Xubuntu
server machine. It contains expression levels for 196 494
splicing variants (57 659 genes) computed in 36 samples:
28 in the normal condition and 8 in the tumoral condition.
For each splicing variant, information related to Ensembl,
Uniprot/Swiss-Prot, Pfam and GO has been integrated as
explained in paragraph ‘Functional Annotation’.
The web interface is implemented through python-CGI programming, HTML and JavaScript. All the graphs are generated through the statistical software R v.2.15.2 and loaded to the web interface through python-CGI. The entity–relationship schema of the database is included in the online DBATE help pages.

Conclusions

Next-generation sequencing technologies revolutionized the analysis of the transcriptome, providing a panoramic view of all the transcriptional activity in a given sample. Although such high-throughput experiments provide an
enormous wealth of data, there are few tools to make order through them. DBATE, freely available at http://bioinformatica.uniroma2.it/DBATE/, provides an integrated resource that can be valuable for the functional inference of whole transcriptome expression analysis experiments by providing pre-computed expression levels and annotations that can be cumbersome to generate for the biomedical scientist.

A semi-automated pipeline was built to process and populate the database with additional data sets as they become available, and will be expanded to more annotation sources and sequencing technologies. The pipeline is based on initial steps of data retrieval, organization and quality checking, done manually by DBATE curators, followed by automated stages of expression estimates and annotations. We initially selected for inclusion in DBATE public data sets from a list retrieved from GEO, including 53 RNA-seq human panels obtained with Illumina technology (GA, GAII, GAIIx or HiSeq). All remaining unprocessed data sets are currently in a queue and will be progressively

Figure 2. Protein interaction network for the ANKRD17, KHSPR, HNRNPK and ANKHD1 genes retrieved from the mentha database and plotted by the mentha browser applet. The mentha database stores manually curated PPIs from five different PPI databases and has been implemented in the DBATE web interface. These four genes have been selected from a complex query search on DBATE to obtain all the splicing variants that encode for protein products containing the KH (K Homology) domain and that are also phosphorylated and contain repeated units. The network includes all primary binding partners of the four genes. Nodes describe genes, and arcs join genes whose protein products are known to physically interact. Nodes corresponding to the query proteins are larger and highlighted with blue circles. Each node is colored according to the expression level of its most expressed splicing variant. Color ranges from red, lower FPKM values; to black, medium expression values; to green, higher expression values. White nodes describe genes for which no splicing variant is expressed in the selected tissue. Protein interaction networks generated by the mentha browser can also be manually expanded and pruned.
added. DBATE updates are planned each semester. Finally, DBATE will be expanded into a web server or web service-based tool for the annotations and characterization of user-submitted RNA-seq panel data.

**Funding**

Associazione Italiana per la Ricerca sul Cancro (AIRC) (IG10298 to M.H.C.). Funding for open access charge: PRIN 2010 (prot. 2010BXYHJS.006 to M.H.C.).

**Conflict of interest.** None declared.

**References**

1. Tress,M.L., Martelli,P.L., Frankish,A. et al (2007) The implications of alternative splicing in the ENCODE protein complement. *Proc. Natl. Acad. Sci. USA*, 104, 5495–5500.
2. Kim,E., Goren,A. and Ast,G. (2008) Alternative splicing: current perspectives. *Bioessays*, 30, 38–47.
3. Pan,Q., Shai,O., Lee,L.J. et al (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnol.*, 28, 503–510.
4. Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12, R22.
5. Ryan,M.C., Cleland,J., Kim,R. et al (2012) SpliceSeq: a resource for analysis and visualization of RNA-Seq data on alternative splicing and its functional impacts. *Bioinformatics*, 28, 2385–2387.
6. Birzele,F., Csaba,G. and Zimmer,R. (2007) Alternative splicing and protein structure evolution. *Nucleic Acids Res.*, 35, 550–558.
7. Stetefeld,J. and Ruegg,M.A. (2005) Structural and functional diversity generated by alternative mRNA splicing. *Trends Biochem. Sci.*, 30, 515–521.
8. Pan,Q., Shai,O., Lee,L.J. et al (2010) Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.*, 12, R22.
9. Nicolae,M., Mangul,S., Mândoiu,I.I. and Zelikovsky,A. (2011) Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Algorithms Mol. Biol.*, 6, 9.
10. Guttmann,M., Garber,M., Levin,J.Z. et al (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnol.*, 28, 503–510.
11. Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12, R22.
12. Ryan,M.C., Cleland,J., Kim,R. et al (2012) SpliceSeq: a resource for analysis and visualization of RNA-Seq data on alternative splicing and its functional impacts. *Bioinformatics*, 28, 2385–2387.
13. Birzele,F., Csaba,G. and Zimmer,R. (2007) Alternative splicing and protein structure evolution. *Nucleic Acids Res.*, 35, 550–558.
14. Stetefeld,J. and Ruegg,M.A. (2005) Structural and functional diversity generated by alternative mRNA splicing. *Trends Biochem. Sci.*, 30, 515–521.
15. Melamud,E. and Moult,J. (2009) Structural implication of splicing stochastics. *Nucleic Acids Res.*, 37, 4862–4872.
16. Leonii,G., Le Pera,L., Ferré,F. et al (2011) Coding potential of the products of alternative splicing in human. *Genome Biol.*, 12, R9.
17. Durbin,R.M., Altshuler,D.L., Durbin,R.M. et al (2010) A map of human genome variation from population-scale sequencing. *Nature*, 467, 1061–1073.
18. Shai,O., Birzele,F., Rosenfeld,M. et al (2012) RNA-Seq Atlas—a reference database for gene expression profiling in normal tissue by next-generation sequencing. *Bioinformatics*, 28, 1184–1185.
19. D’Antonio,M., Bonizzoni,P. et al (2011) AS-PiDB: a database of annotated transcript and protein variants generated by alternative splicing. *Nucleic Acids Res.*, 39, D80–D85.
20. Birzele,F., Köffner,R., Meier,F. et al (2008) ProSAS: a database for analyzing alternative splicing in the context of protein structures. *Nucleic Acids Res.*, 36, D63–D68.
21. Shionyu,M., Takahashi,K. and Go,M. (2012) AS-DEST: a functional annotation tool for putative proteins encoded by alternatively spliced transcripts. *Bioinformatics*, 28, 2076–2077.
22. Floris,M., Raimondo,D., Leoni,G. et al (2011) MAISTAS: a tool for automatic structural evaluation of alternative splicing products. *Bioinformatics*, 27, 1625–1629.
23. Flicek,P., Amode,M.R., Barrell,D. et al (2011) Ensembl 2012. *Nucleic Acids Res.*, 40, D84–D90.
24. Martelli,P.L., Floris,M., Raimondo,D. et al (2011) Sex-specific and tissue-specific alternative splicing in human genome variation from population-scale sequencing. *Nature Genet.*, 43, 213–220.
25. Kim,N., Alekseyenko,A.V., Roy,M. and Lee,C. (2007) The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species. *Nucleic Acids Res.*, 35, D93–D98.
26. Hon,G.C., Hawkins,R.D., Caballero,O.L. et al (2011) Improving RNA-Seq data alignment accuracy using multi-exonic splicing events. *Bioinformatics*, 27, D1005–D1010.
27. Wang,E.T., Sandberg,R., Luo,S. et al (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, 455, 470–476.
28. Blekhman,R., Marioni,J.C., Zumbo,P. et al (2010) Sex-specific and lineage-specific alternative splicing in primates. *Genome Res.*, 20, 180–189.
29. Wang,E.T., Sandberg,R., Luo,S. et al (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, 455, 470–476.
30. Wang,E.T., Sandberg,R., Luo,S. et al (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, 455, 470–476.
36. Mercer,T.R., Neph,S., Dinger,M.E. et al. (2011) The human mitochondrial transcriptome. Cell, 146, 645–658.
37. Reich,A., Klatsky,P., Carson,S. and Wessel,G. (2011) The transcriptome of a human polar body accurately reflects its sibling oocyte. J. Biol. Chem., 286, 40743–40749.
38. Chen,L.Y., Wei,K.C., Huang,A.C. et al. (2012) RNASEQR—a streamlined and accurate RNA-seq sequence analysis program. Nucleic Acids Res., 40, e42.
39. Mullokandov,G., Baccarini,A., Ruoz,A. et al. (2012) High-throughput assessment of microRNA activity and function using microRNA sensor and decoy libraries. Nat. Methods, 9, 840–846.
40. Bert,S.A., Robinson,M.D., Strbenac,D. et al. (2013) Regional activation of the cancer genome by long-range epigenetic remodeling. Cancer cell, 23, 9–22.
41. Miao,F., Chen,Z., Zhang,L. et al. (2013) RNA-sequencing analysis of high glucose treated monocytes reveals novel transcriptome signatures and associated epigenetic profiles. Physiol. Genomics, 45, 287–299.
42. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol., 10, R25.
43. Trapnell,C., Pachter,L. and Salzberg,S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 25, 1105–1111.
44. Trapnell,C., Roberts,A., Goff,L. et al. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc., 7, 562–578.
45. Kim,H., Bi,Y., Pa,l2. et al. (2011) IsoformEx: isoform level gene expression estimation using weighted non-negative least squares from RNA-Seq data. BMC Bioinformatics, 12, 305.
46. Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics, 12, 323.
47. Glaus,P., Honkela,A. and Rattray,M. (2012) Identifying differentially expressed transcripts from RNA-seq data with biological variation. Bioinformatics, 28, 1721–1728.
48. Du,J., Leng,J., Habegger,L. et al. (2012) IQSeq: integrated isoform quantification analysis based on next-generation sequencing. PLoS One, 7, e29175.
49. Mercer,T.R., Dinger,M.E. and Mattick,J.S. (2009) Long non-coding RNAs: insights into functions. Nat. Rev. Genet., 10, 155–159.
50. Ponting,C.P., Oliver,P.L. and Reik,W. (2009) Evolution and functions of long noncoding RNAs. Cell, 136, 629–641.
51. Prensner,J.R. and Chinnaiyan,A.M. (2011) The emergence of IncRNAs in cancer biology. Cancer Discov., 1, 391–407.
52. Ambros,V. (2004) The functions of animal microRNAs. Nature, 431, 350–355.
53. Kerrien,S., Aranda,B., Breuza,L. et al. (2012) The IntAct molecular interaction database in 2012. Nucleic Acids Res., 40, D841–D846.
54. Licausa,L., Brugnoli,I., Peluso,D. et al. (2012) MINT, the molecular interaction database: 2012 update. Nucleic Acids Res., 40, D857–D861.
55. Salwinski,L., Miller,C.S., Smith,A.J. et al. (2004) The database of interacting proteins: 2004 update. Nucleic Acids Res., 32, D449–D451.
56. Stark,C., Breitkreutz,B.J., Chatr-Aryamontri,A. et al. (2011) The BioGRID interaction database: 2011 update. Nucleic Acids Res., 39, D698–D704.
57. Chautard,E., Fatoux-Ardore,M., Ballut,L. et al. (2011) MatrixDB, the extracellular matrix interaction database. Nucleic Acids Res., 39, D508–D516.
58. Garcia-Mayoral,M.F., Diaz-Moreno,L., Hollingworth,D. and Ramos,A. (2008) The sequence selectivity of KSRP explains its flexibility in the recognition of the RNA targets. Nucleic Acids Res., 36, 5290–5296.
59. Iijima,T., Wu,K., Witte,H. et al. (2011) SAM68 regulates neuronal activity-dependent alternative splicing of neurexin-1. Cell, 147, 1601–1614.
60. Smith,R.K., Carroll,P.M., Allard,J.D. and Simon,M.A. (2002) MASK, a large ankyrin repeat and KH domain-containing protein involved in Drosophila receptor tyrosine kinase signaling. Development, 129, 71–82.
61. Traina,F., Favaro,P.M., Medina Sde,S. et al. (2006) ANKHD1, ankyrin repeat and KH domain-containing protein involved in K562 cells. Biochim. Biophys. Acta, 1762, 828–834.
62. Diaz-Moreno,L., Hollingworth,D., Frenkel,T.A. et al. (2009) Phosphorylation-mediated unfolding of a KH domain regulates KSRP localization via 14-3-3 binding. Nat. Struct. Mol. Biol., 16, 238–246.
63. Miles,M.C., Janket,M.L., Wheeler,E.D. et al. (2005) Molecular and functional characterization of a novel splice variant of ANKHD1 that lacks the KH domain and its role in cell survival and apoptosis. FEBS J., 272, 4091–4102.