OXFORD

Briefings in Bioinformatics, 18(1), 2017, 28–36

doi: 10.1093/bib/bbv117
Advance Access Publication Date: 27 January 2016
Paper

MTD: a mammalian transcriptomic database to explore gene expression and regulation

Xin Sheng*, Jiayan Wu*, Qianqian Sun, Xue Li, Feng Xian, Manman Sun, Wan Fang, Meili Chen, Jun Yu and Jingfa Xiao

Corresponding author. Jingfa Xiao, CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, 100101, China. Tel.: +86-10-84097443; E-mail: xiaojingfa@big.ac.cn; Jun Yu, CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, 100101, China. Tel.: +86-10-84097898; E-mail: junyu@big.ac.cn
*These authors contributed equally to this work.

Abstract

A systematic transcriptome survey is essential for the characterization and comprehension of the molecular basis underlying phenotypic variations. Recently developed RNA-seq methodology has facilitated efficient data acquisition and information mining of transcriptomes in multiple tissues/cell lines. Current mammalian transcriptomic databases are either tissue-specific or species-specific, and they lack in-depth comparative features across tissues and species. Here, we present a mammalian transcriptomic database (MTD) that is focused on mammalian transcriptomes, and the current version contains data from humans, mice, rats and pigs. Regarding the core features, the MTD browses genes based on their neighboring genomic coordinates or joint KEGG pathway and provides expression information on exons, transcripts and genes by integrating them into a genome browser. We developed a novel nomenclature for each transcript that considers its genomic position and transcriptional features. The MTD allows a flexible search of genes or isoforms with user-defined transcriptional characteristics and provides both table-based descriptions and associated visualizations. To elucidate the dynamics of gene expression regulation, the MTD also enables comparative transcriptomic analysis in both intraspecies and interspecies manner. The MTD thus constitutes a valuable resource for transcriptomic and evolutionary studies. The MTD is freely accessible at http://mtd.cbi.ac.cn.

Xin Sheng is a PhD candidate at the CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics and University of the Chinese Academy of Sciences, Beijing, China. She has been working in the field of mammalian transcriptomic study.
Jiayan Wu is an associate professor at the CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China. Her research interest is focused on comparative genomics and bioinformatics.
Qianqian Sun is a master candidate at the CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics and University of the Chinese Academy of Sciences, Beijing, China.
Xue Li is a master candidate at the CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics and University of the Chinese Academy of Sciences, Beijing, China.
Feng Xian is a PhD candidate at the CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics and University of the Chinese Academy of Sciences, Beijing, China.
Manman Sun is an undergraduate student at Hunan Agricultural University, Hunan, China.
Wan Fang is a master candidate at the CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics and University of the Chinese Academy of Sciences, Beijing, China.
Meili Chen is a PhD at the CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China.
Jun Yu is a professor at the CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China. His research interest is focused on transcriptomics and bioinformatics.
Jingfa Xiao is a professor at the CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China. His research interest is focused on transcriptomics and bioinformatics.
Submitted: 2 December 2015; **Received (in revised form):** 14 December 2015

Key words: mammalian transcriptomic database; gene expression and regulation; RNA-seq

Introduction

Transcriptomes of cells and organs, which connect genome information to gene operations, are both coordinated and intricate in terms of functional pathways and spatiotemporal organizations. Mammalian transcriptomes are most relevant to human diseases and have been studied in depth [1]. An obvious stratification of the mammalian transcriptomes is to split the genes into housekeeping (HK) and tissue-specific (TS) genes [2]; both classes exhibit distinct features, such as gene structure [3], mutation rate [4], chromosomal organization [5] and replication timing [6]. Other definitions and ideas about HK genes have also been proposed [7]. Furthermore, alternative transcripts complicate the definition of transcriptomes, including alternative transcription initiation, splicing and polyadenylation [1]. Recent studies have suggested that the expression of TS genes is largely conserved, whereas alternative splicing tends to be lineage-specific among mammals [8–10]; however, this point remains controversial [11]. Nevertheless, comprehensive transcriptome surveys and a unified database are always essential for future repeated scrutiny as the data are being accumulated.

Recent next-generation RNA sequencing (RNA-seq) has made it possible to extensively examine transcriptomic regulation in different tissues and implement a more concrete description of the transcriptome [12]. The overwhelming volume of public RNA-seq data requires effective data mining and integration. Current mammalian transcriptomic databases are either TS or species-specific. For example, BloodExpress [13] focuses on murine blood cell type expression profiles in distinct differentiation stages. Allen Brain Atlas [14] is built on gene expression data, connectivity data and neuroanatomical information for the adult and developing brain in mouse and primate. Wikicell [15] is a wiki-based database that provides a transcriptome model based on a human taxonomy graph. EMAGE [16] harbors text-based descriptions of gene expression and spatial maps of the gene expression patterns of mouse embryos. NHPRTR [17] is a community resource of comprehensive reference transcriptomes from multiple primates. Finally, Expression Atlas [18] is a comprehensive database, and it contains information on gene, protein and splice variant expression in different cell types, organism parts, developmental stages, diseases and other biological and experimental conditions. Although these existing databases offer abundant transcriptomic information in various developmental stages or under different physiological conditions in multiple species, an in-depth investigation of transcriptional features and a parallel comparison of the transcriptomes across species are still lacking.

Unique among these related databases, we present our developed mammalian transcriptomic database (MTD), which is a mammalian transcriptomic database focused on characterization and comparative analyses of RNA-seq data in most available tissues/cell lines of humans, mice, rats and pigs. First, the MTD allows browsing genes by their neighboring genomic coordinates or their joint KEGG pathway and provides detailed expression characteristics of exon, transcript and gene levels by embedding a powerful genome browser; that is, GBrowse. Second, based on transcriptional features and the genomic position of each transcript, we developed novel nomenclature that allows flexible searching of genes or isoforms with user-specified transcriptional characteristics. Third, the MTD also provides a critical function to elucidate the dynamics of gene expression regulation across tissues/cell lines and species, which is a comparative transcriptomic analysis that can be performed intraspecies and interspecies. Moreover, to provide intuitive results, the MTD presents queried transcriptomic information by combining a table-based description with an associated visualization. The MTD thus constitutes a unique and valuable new resource to promote related transcriptomic and evolutionary studies.

Methods and materials

Raw RNA-seq samples filtering and preprocessing

Figure 1 shows the pipeline of data processing and database construction. All raw RNA-seq data and the corresponding experiment metadata (.xml format) of four model mammals (human, pig, rat and mouse) were collected from healthy samples without any special treatment in Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra) by considering the five criteria mentioned as follows: (1) only selecting samples under normal conditions and removing tissues/cell lines with special treatments to avoid different regulation mechanisms under diverse physiological conditions, (2) removing all mixed tissue/cell line samples that might mask the real differences among samples, (3) filtering severe unsaturated data sets that might increase the false-negative rate of gene expression and the HK genes, (4) selecting the most saturated and highest sequencing quality experiment as representative data for each tissue/cell line and (5) selecting samples from the Illumina Genome Analyzer and Illumina HiSeq 2000/2500 to reduce the differences resulting from the use of different sequencing instruments (Table 1). Then, we filtered raw reads for selecting samples according to the six criteria mentioned as follows: (1) filtering out adaptor reads, (2) truncating reads with more than 2% 'N' bases, (3) truncating reads with low quality (below 20) over 20% of the length, (4) truncating reads with low quality (below 13) over 10% of the length, (5) truncating reads with a sequence quality of less than 20 and (6) after truncating, any human and mouse reads shorter than 50 bp (36 bp for rat and pig because of the poor quality of the relevant RNA-seq resources) were filtered out.

Read alignment and transcript assembly

All the remaining reads of each sample were mapped to their respective genomes (hg19, rn4, Sus scrofa10.2 and mm 10) using Tophat [19] version 2.0.9. Additionally, we used Bedtools (http://bedtools.readthedocs.org/en/latest/) and our Perl scripts to perform saturation analysis for each tissue/cell line. Cufflinks [20] version 2.1.1 was used to assemble transcripts and calculate the read density (shown as reads per kilobase per million mapped reads (RPKM) value) [21] of each gene or isoform by the read number that was uniquely mapped. As for each exon, we added all RPKMs of isoforms that contained that exon as its RPKM. To provide a comprehensive reference data set to users, we identified HK genes and HK isoforms by limiting RPKM >0 in all the representative experiments of the tissues/cell lines with saturated data. Moreover, we calculated the coefficients of variation (CVs) of HK genes and/or HK isoforms and percent spliced-in (PSIs) of exons and extracted the transcriptional features of genes and isoforms with our Perl scripts.

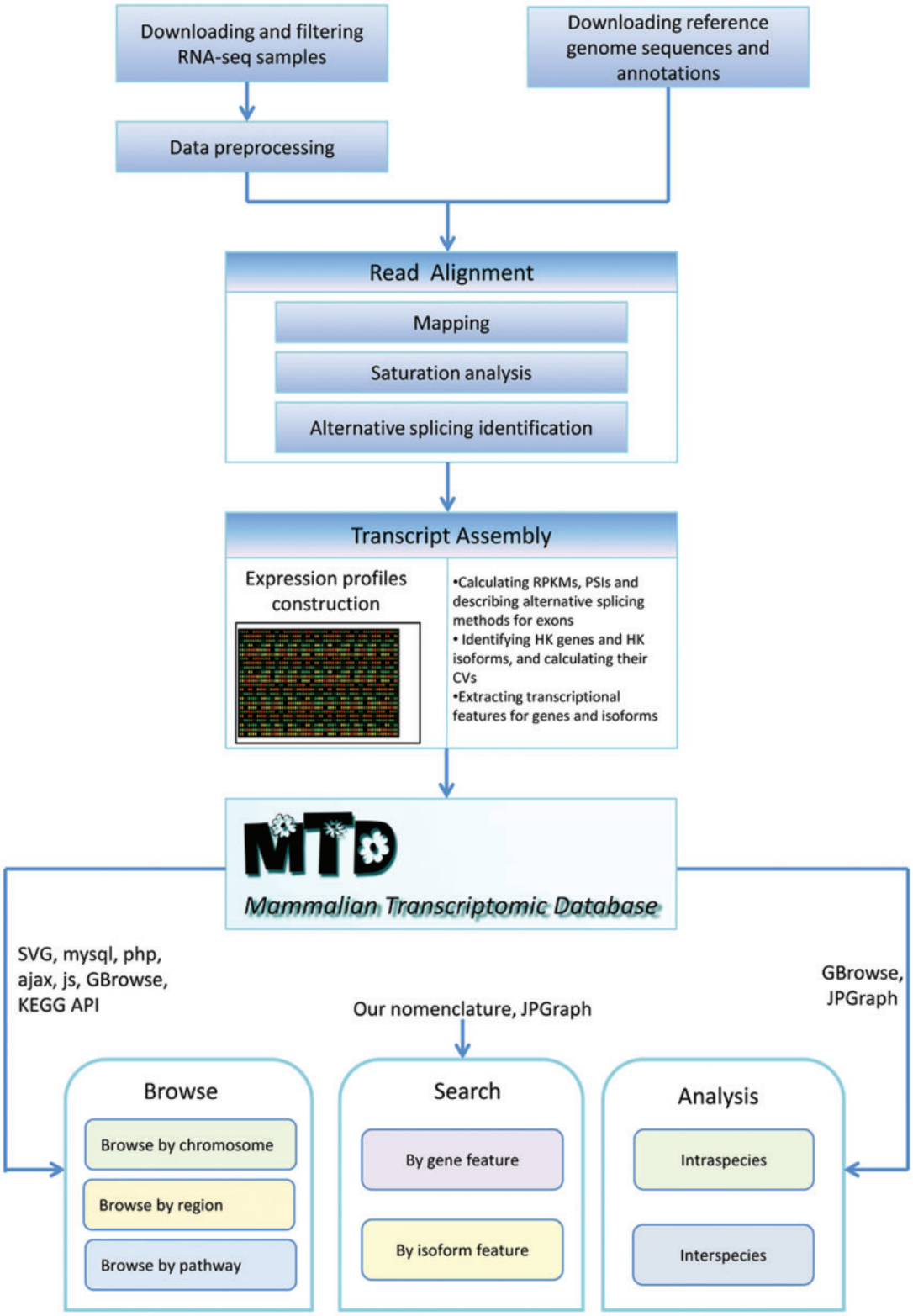

Figure 1. The pipeline of data processing and database construction.

Notably, we tried to retrieve all accessible data in SRA of the four mammals. For the tissue/cell line with multiple experiments, we selected the best one as representative experiment by considering sequencing quality, saturation degree and experiment design. However, the other data were also exhibited in MTD to serve as a supplement. The data processes of read alignment, saturation analysis and transcript assembly for all experiments were performed independently. As for the experiment with technical replicates, we ordered the replicates by sequencing quality and saturation degree, and then chose the best one as

Table 1. The data statistics of the MTD, which contains the total amount of available data, selected data, the saturated data and the covered experiments, projects and tissues/cell lines of these selected data

Species	Available data	Selected data	Saturated data	Covered experiments	Covered projects	Covered tissues/cell lines
Homo sapiens	228	83	72	75	24	31
Sus scrofa	148	27	18	27	11	13
Rattus norvegicus	180	36	36	36	10	14
Mus musculus	1059	108	88	86	40	44

the major source for expression results description of this experiment and the others as supplements. The data of the experiments in each tissue/cell line are displayed individually in MTD.

Implements

The Web site was written in the LAMP (Linux, Apache, MySQL and PHP) environment. We arranged our transcriptomic data to fit into our designed relational database schema and stored the data in the MySQL database (http://www.mysql.org; a free and popular relational database management system; version 5.1.69). The MTD was built using HyperText Markup Language (HTML and HTML5), HyperText Preprocessor (PHP) (http://php.net; a widely used general-purpose scripting language; version 5.2.17), JavaScript, jQuery, Cascading Style Sheets (CSS, CSS3), Scalable Vector Graphics (SVG) and Asynchronous JavaScript and XML (AJAX). To make pages appealing and user-friendly, we used JavaScript, CSS and CSS3. JavaScript, jQuery and AJAX were used to add actions and allow for asynchronously refreshing pages. We also drew chromosomal ideograms with SVG to ensure high resolution and dynamic reflection. The structure images and coverage plots of genes were supported by embedding GBrowse [22], and KEGG [23] pathway graphs were provided by KEGG API (http://www.kegg.jp/kegg/rest/). JPGraph (http://jpgraph.net) was also used to offer an overview of gene expression levels across tissues/cell lines for each expressed gene/isoform. The MTD is freely available at http://mtd.cbi.ac.cn.

Results

Database content and usage

The MTD can be accessed through a user-friendly Web interface. Online documentation is provided to help users access the database. We have collected 83, 27, 36 and 108 RNA-seq data of humans, pigs, rats and mice, respectively (covering 31, 13, 14 and 44 tissues/cell lines, respectively), in the MTD. The MTD was designed with five main functionalities for data retrieval: Browse, Search, Analysis, Visualization and Download.

Browse

We use three methods to guide users to browse for a specific set of genes. First, we provide 'browse by chromosome' functions for humans and mice (Figure 2A) because of their well-annotated and assembled reference genomes. The MTD enables browsing for both gene expression levels and read coverage information across tissues/cell lines/experiments with neighboring genomic coordinates or in a specific chromosomal cytoband. Each gene symbol links to the detailed transcriptomic information page of its isoforms, and each RefSeq ID of isoform links to a further transcriptomic feature page of its exons. This feature also exists for the other function pages. Second, to facilitate users exporting structure and coverage plots of genes, searching their interested genome regions in a specific data source and setting their investigational tracks more flexibly, we embedded GBrowse (Figure 2B). By inputting a chromosomal region formatted as 'chromosome:start-end' or a gene symbol or RefSeq ID, users can browse read coverage information in a specific tissue/cell line/experiment of their chromosomal region of interest. Moreover, additional tracks, including exon, intron, DNA, restriction sites and 6-frame translation, can be set by users for browsing. All the resulting images (read coverage plots and gene structure plots in the queried chromosomal region) can be exported. Third, when combined with KEGG API [23], the MTD can browse gene expression levels based on their joint KEGG pathway in selected tissues/cell lines/experiments (Figure 2C), each image of KEGG pathway links to the details description page of KEGG and each resulting table is sortable, which make it easy to find genes with high and/or low expression levels in the pathway of interest.

Search and analysis

For aspects of both genes and isoforms, it is important to know their transcriptional features. Here, based on the transcriptional characteristics of genes and isoforms, users can flexibly search by limiting different combinations of standards of transcriptional features for genes or isoforms. The MTD enables an online real-time graphing histogram (by JPGraph) that provides an overview of the gene expression levels across tissues/cell lines for each expressed gene/isoform (Figure 3A). It also allows for the identification of HK genes or isoforms by user-defined cutoffs of RPKM or CV online. Genes or isoforms with specific transcriptional features in a particular tissue/cell line can also be obtained (Figure 3B). Moreover, for each well-characterized gene, a KEGG page containing the comprehensive information will be linked. When studying a specific tissue/cell line, users can obtain its gene expression profile with the approved cutoffs for RPKM or CV. In addition, the related pages recording the raw experimental conditions for each tissue/cell line in the SRA can be linked below the corresponding query result table. Regarding isoforms, by considering the genomic position and transcriptional features (including alternative splicing types, transcriptional directions and expression levels) of each transcript, we gave each isoform a unique name. Based on this nomenclature, the MTD makes it convenient for users to search for isoforms with specific transcriptional features (Figure 3C and D). The detailed introduction for this nomenclature of locus-specific transcripts can be found on our 'FAQ' page. Notably, all of the transcriptional features in this function module (Search) are based on the transcriptomic results from the representative experiment for each tissue/cell line.

Although it is useful to obtain transcripts or genes with transcriptional attributes of interest, it is also valuable to perform comparative transcriptomic analyses that are both intraspecies and interspecies, which may help researchers elucidate the dynamics of gene expression regulation across tissues and species. An intraspecies interface (Figure 4A), which allows comparison of transcriptomes across tissues or cell lines on gene, transcript and exon levels, is provided in the MTD. An

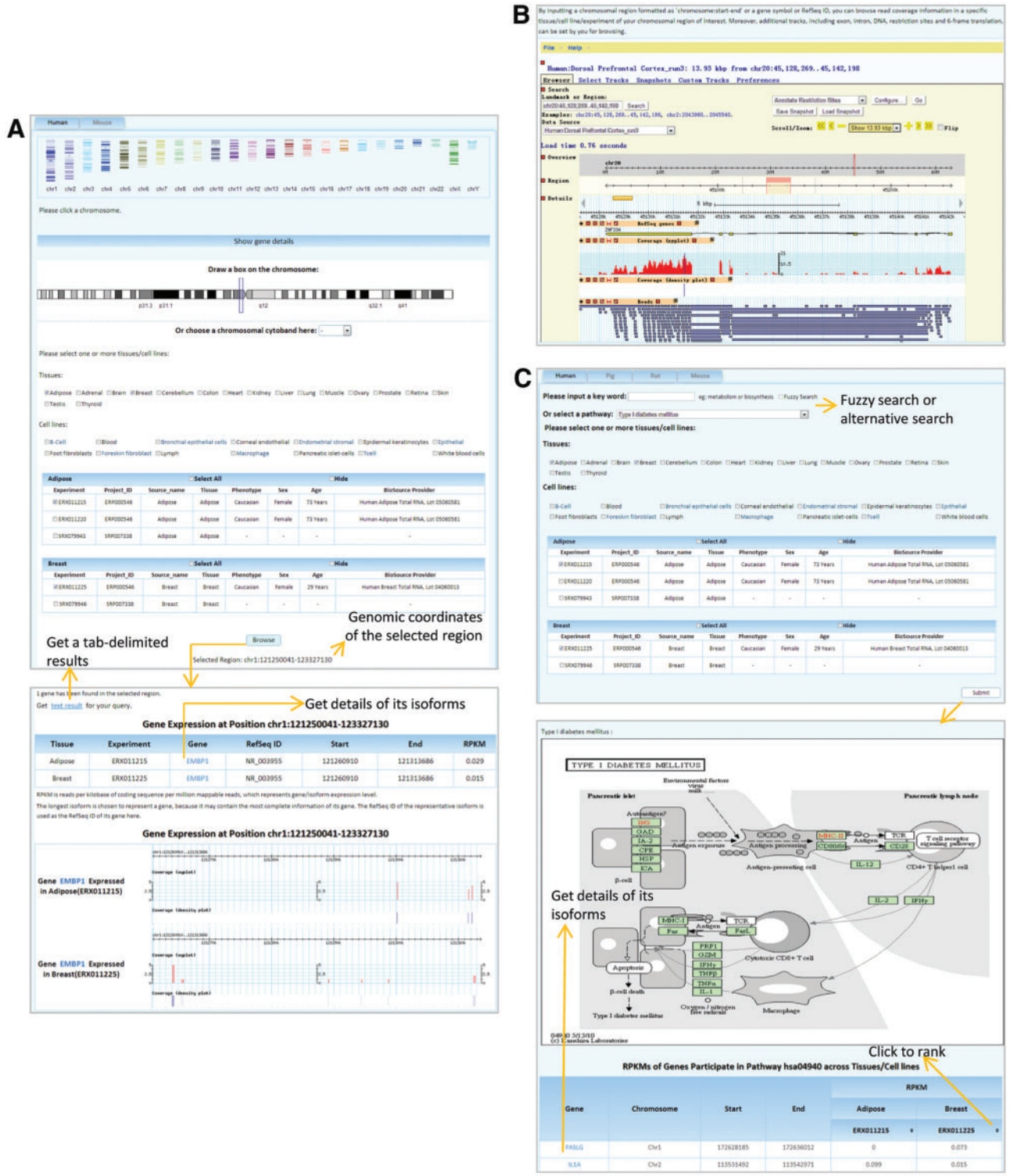

Figure 2. Representative screen shots of the Browse functionalities. (**A**) Browse by chromosome. (**B**) Browse by region. (**C**) Browse by pathway.

interspecies interface allows comparisons of the transcriptomic details of homologous genes in physiologically equivalent tissues across species (Figure 4B). The homologous genes were identified according to the corresponding amino acid sequences of protein-coding genes and the nucleic acid sequences of non-coding genes by OrthoMCL [24] version 2.0.9 with the default cutoff of E value. Users need to enter a gene symbol or a RefSeq ID and choose which species the searching gene originates from. In response to this query, the MTD provides a histogram that vividly and comprehensively presents the expression levels of both the searched gene and its orthologous genes in the representative experiment of each comparative tissue/cell line for the compared species. This histogram gives an overview of the comparison results and the detailed analytical information on

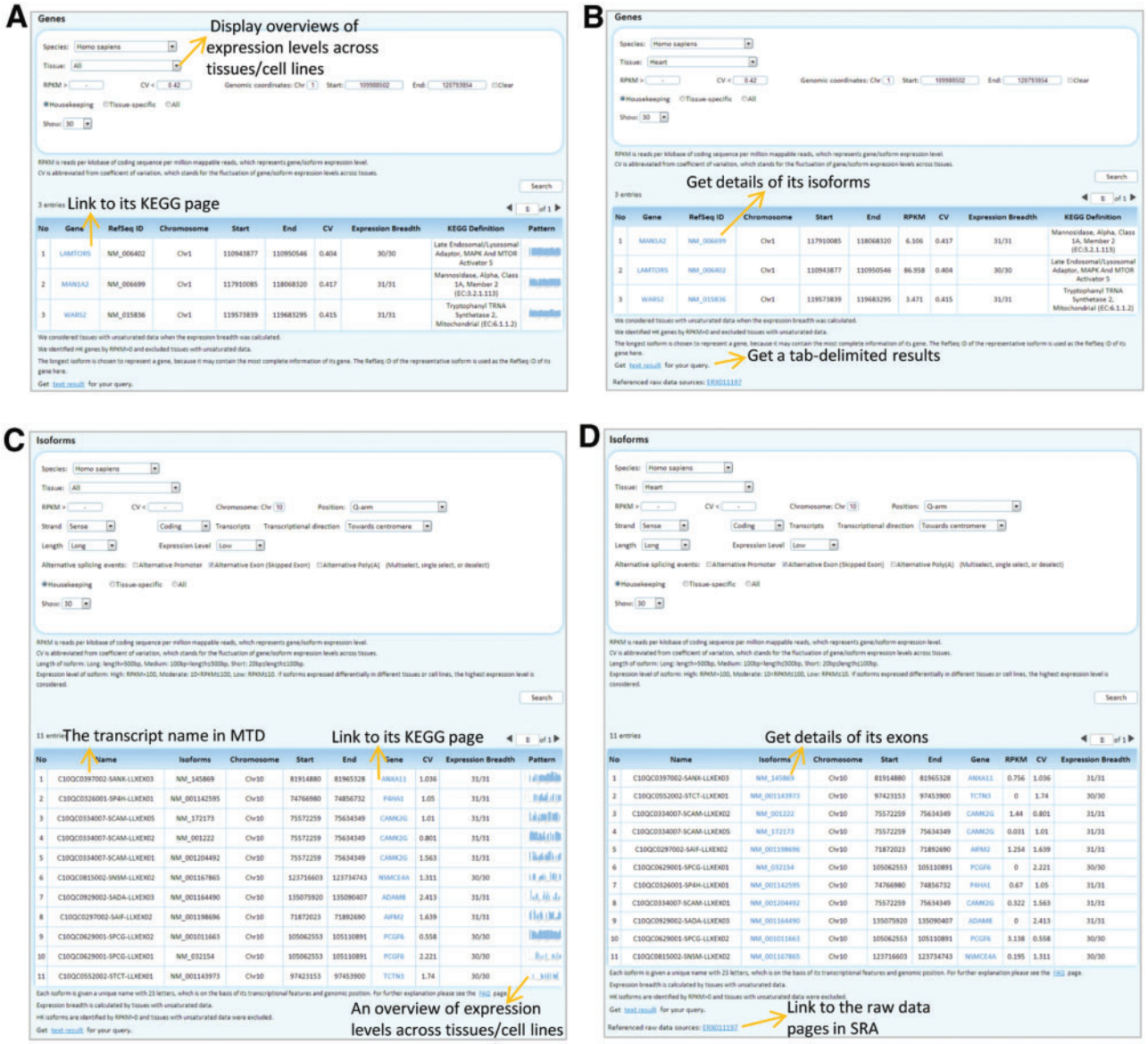

Figure 3. Representative screen shots of the Search functionalities. For genes, (**A**) view overviews of the expression levels across tissues/cell lines. (**B**) Search genes with specific transcriptional features in a specific tissue/cell line. For isoforms, (**C**) view overviews of the expression levels across tissues/cell lines. (**D**) Search isoform with specific transcriptional features in a specific tissue/cell line.

gene, transcript and exon levels in all the related experiments, which can be accessed for each tissue/cell line by clicking on the 'show details' button at the bottom. In addition, we intuitively present the queried results using a text-based table combined with a visualization of the gene structure and its read coverage plot in each tissue/cell line/experiment. In conclusion, the MTD allows users to inspect transcriptome information not only from the gene level to the exon level but also according to an overview of the gene expression levels across all comparative tissues/cell lines or species.

Visualization and download

GBrowse is embedded to intuitively display the results. The queried results can be visualized at three different levels, namely, gene, isoform and exon, which show different alternative splicing patterns of different isoforms and the distribution of expression across the different exons and isoforms of a given gene. In addition, the MTD provides gene and isoform expression profiles in the representative experiments of all collected tissues/cell lines; basic information of genes, isoforms and exons (extracted from reference annotations); isoform nomenclature; and the HK genes, HK isoforms and orthologous genes of four mammals on the 'Download' page. Each queried table can be downloaded as a tab-delimited text for further research.

Discussion and future directions

The MTD focuses on characterization and comparative analysis of RNA-seq data in most available tissues of humans, mice, rats and pigs, and can be accessed through a user-friendly Web interface (http://mtd.cbi.ac.cn). Unlike the extant databases, the 'Browse' interfaces allow the unique and useful investigation of transcriptomes for two collections of genes, including genes with

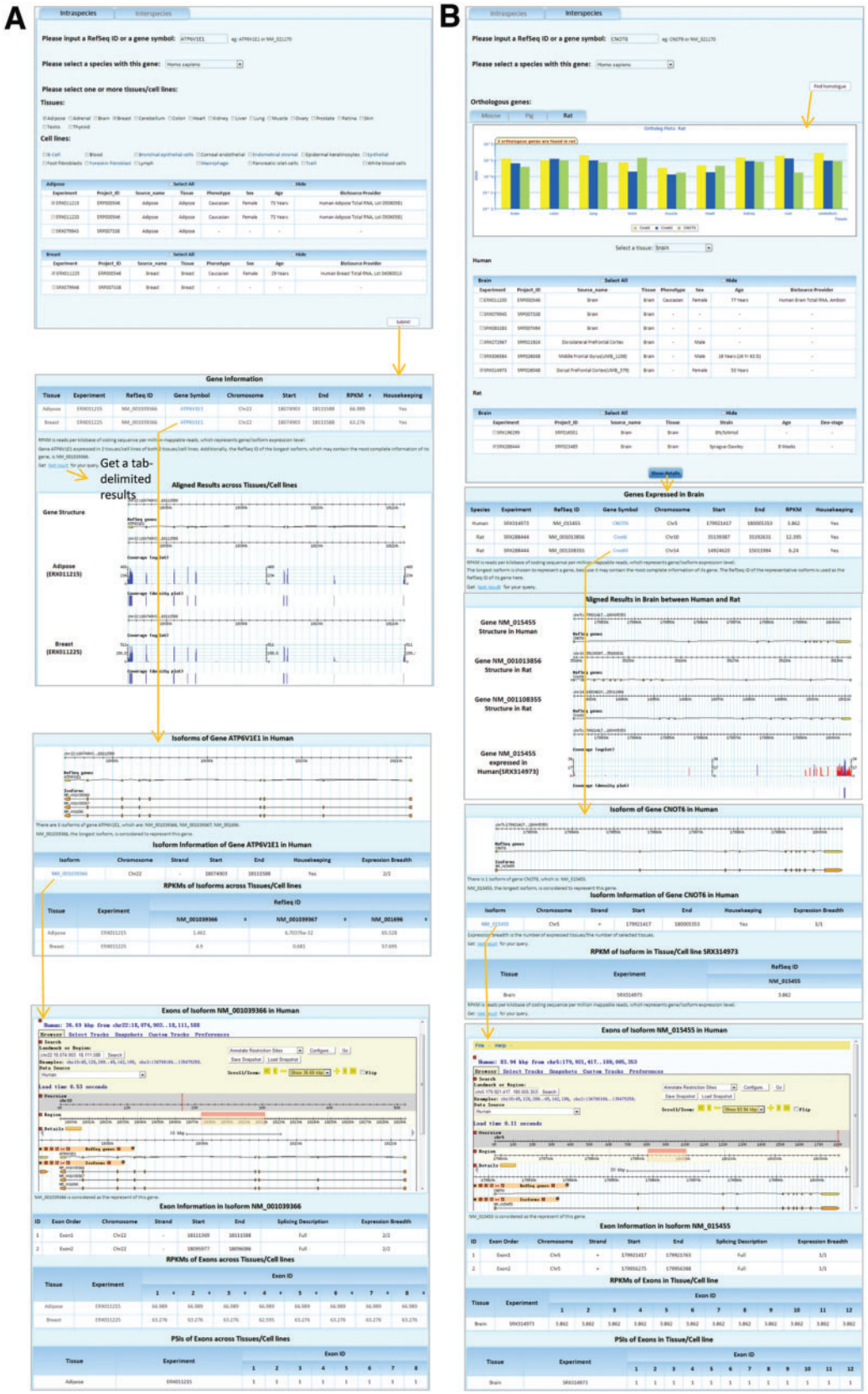

Figure 4. Representative screen shots of the Analysis functionalities. (**A**) An intraspecies interface allows the comparison of transcriptomes across tissues or cell lines on gene, transcript and exon levels. (**B**) An interspecies interface allows comparison of the transcriptomic details of homologous genes in physiologically equivalent tissues/cell lines across species.

neighboring genomic coordinates and genes active in the same biological pathway. By embedding GBrowse, the read distributions across the different exons and transcripts of a given gene can be intuitively presented [18]. In addition, the flexibility of filtering by the transcriptional features of genes and isoforms enhances the capability of users for discovering transcriptomic stories for genes or isoforms with specific transcriptional characteristics, such as HK genes [7, 25], expression profiles of tissues/cell lines [8–11] and isoforms undergoing an 'exon skipped' alternative splicing event [1, 8, 10]. A comparative transcriptomic analysis that is both intraspecies and interspecies provides a critical function to elucidate the dynamics of gene expression regulation across tissues and species [8–10]. Moreover, the MTD affords the detailed expression characteristics of exon, transcript and gene levels [7]. Accordingly, the MTD satisfies a previously unmet need for related transcriptomic and evolutionary research by allowing for deeper and multiperspective investigation of the complicated and dynamic transcriptome.

Aside from adding more tissue/cell line types for existing species, we will recruit other mammal species when their reference genomes meet the standard. In addition, we will also integrate other data types, such as single-cell RNA-seq and epigenomic ChIP-seq data, to support new expression patterns [11] and mechanisms.

Key Points

- To provide a valuable resource for transcriptomic and evolutionary studies, we developed a mammalian transcriptomic database with integration of 83, 27, 36 and 108 RNA-seq data for humans, pigs, rats and mice.
- The MTD browses genes based on their neighboring genomic coordinates or joint KEGG pathway and provides expression information on exons, transcripts and genes to facilitate users to understand the gene expression and regulation from various perspectives.
- The MTD allows a flexible search of genes or isoforms with user-defined transcriptional characteristics and provides both table-based descriptions and associated visualizations.
- The MTD enables comparative transcriptomic analysis in both intraspecies and interspecies manners to elucidate the dynamics of gene expression regulation.

Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Acknowledgements

The authors thank Dr Songnian Hu for valuable discussions on this work and Xiaoman Bi for reporting bugs and sending comments.

Funding

This work was supported by National Programs for High Technology Research and Development [863 Program; 2015AA020108, 2012AA020409 to J.X.]; National Basic Research and Development Program [973 Program; 2011CB944101 to J.Y.]; National Natural Science Foundation of China [31271386 to J.X.]; and funding for open access charge: National Programs for High Technology Research and Development [2012AA020409].

References

1. Adams J. Transcriptome: connecting the genome to gene function. *Nat Educ* 2008;**1**:195.
2. Chen M, Xiao J, Zhang Z, *et al*. Identification of human HK genes and gene expression regulation study in cancer from transcriptomics data analysis. *PloS One* 2013;**8**:e54082.
3. Zhu J, He F, Wang D, *et al*. A novel role for minimal introns: routing mrnas to the cytosol. *PloS One* 2010;**5**:e10144.
4. Cui P, Lin Q, Ding F, *et al*. The transcript-centric mutations in human genomes. *Genomics Proteomics Bioinformatics* 2012;**10**:11–22.
5. Cui P, Liu W, Zhao Y, *et al*. The association between H3K4me3 and antisense transcription. *Genomics Proteomics Bioinformatics* 2012;**10**:74–81.
6. Cui P, Ding F, Lin Q, *et al*. Distinct contributions of replication and transcription to mutation rate variation of human genomes. *Genomics Proteomics Bioinformatics* 2012;**10**:4–10.
7. Eisenberg E, Levanon EY. Human housekeeping genes, revisited. *Trends Genet* 2013;**29**:569–74.
8. Barbosa-Morais NL, Irimia M, Pan Q, *et al*. The evolutionary landscape of alternative splicing in vertebrate species. *Science* 2012;**338**:1587–93.
9. Merkin J, Russell C, Chen P, *et al*. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science* 2012;**338**:1593–9.
10. Reyes A, Anders S, Weatheritt RJ, *et al*. Drift and conservation of differential exon usage across tissues in primate species. *Proc Natl Acad Sci USA* 2013;**110**:15377–82.
11. Lin S, Lin Y, Nery JR, *et al*. Comparison of the transcriptional landscapes between human and mouse tissues. *Proc Natl Acad Sci USA* 2014;**111**:17224–9.
12. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**:57–63.
13. Miranda-Saavedra D, De S, Trotter MW, *et al*. BloodExpress: a database of gene expression in mouse haematopoiesis. *Nucleic Acids Res* 2009;**37**:D873–9.
14. Sunkin SM, Ng L, Lau C, *et al*. Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Res* 2013;**41**:D996–1008.
15. Zhao D, Wu J, Zhou Y, *et al*. WikiCell: a unified resource platform for human transcriptomics research. *OMICS* 2012;**16**:357–62.
16. Richardson L, Venkataraman S, Stevenson P, *et al*. EMAGE mouse embryo spatial gene expression database: 2014 update. *Nucleic Acids Res* 2014;**42**:D835–44.
17. Peng XX, Thierry-Mieg J, Thierry-Mieg D, *et al*. Tissue-specific transcriptome sequencing analysis expands the non-human primate reference transcriptome resource (NHPRTR). *Nucleic Acids Res* 2015;**43**:D737–42.
18. Petryszak R, Burdett T, Fiorelli B, *et al*. Expression Atlas update-a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res* 2014;**42**:D926–32.
19. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009;**25**:1105–11.
20. Trapnell C, Roberts A, Goff L, *et al*. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 2012;**7**:562–78.
21. Mortazavi A, Williams BA, McCue K, *et al*. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008;**5**:621–8.

22. Stein LD, Mungall C, Shu SQ, *et al*. The generic genome browser: a building block for a model organism system database. *Genome Res* 2002;**12**:1599–610.
23. Kanehisa M, Goto S, Sato Y, *et al*. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* 2014;**42**:D199–205.
24. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 2003; **13**:2178–89.
25. Zhu J, He FH, Song SH, *et al*. How many human genes can be defined as housekeeping with current expression data? *BMC Genomics* 2008;**9**:11.