D86–D90 Nucleic Acids Research, 2010, Vol. 38, Database issue Published online 7 December 2009
doi:10.1093/nar/gkp984

H-DBAS: human-transcriptome database for alternative splicing: update 2010

Jun-ichi Takeda[1,2], Yutaka Suzuki[2], Ryuichi Sakate[1], Yoshiharu Sato[1], Takashi Gojobori[1,3], Tadashi Imanishi[1,*] and Sumio Sugano[2]

[1]Integrated Database and Systems Biology Team, Biomedicinal Information Research Center National Institute of Advanced Industrial Science and Technology, AIST Bio-IT Research Bldg. Aomi 2-4-7, Koto-ku, Tokyo 135-0064, [2]Department of Medical Genome Sciences, Graduate School of Frontier Sciences, the University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8562 and [3]Center for Information Biology and DDBJ, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan

Received September 15, 2009; Revised October 13, 2009; Accepted October 15, 2009

ABSTRACT

H-DBAS (http://h-invitational.jp/h-dbas/) is a specialized database for human alternative splicing (AS) based on H-Invitational full-length cDNAs. In this update, for better annotations of AS events, we correlated RNA-Seq tag information to the AS exons and splice junctions. We generated a total of 148 376 598 RNA-Seq tags from RNAs extracted from cytoplasmic, nuclear and polysome fractions. Analysis of the RNA-Seq tags allowed us to identify 90 900 exons that are very likely to be used for protein synthesis. On the other hand, 254 AS junctions of human RefSeq transcripts are unique to nuclear RNA and may not have any translational consequences. We also present a new comparative genomics viewer so that users can empirically understand the evolutionary turnover of AS. With the unique experimental data closely connected with intensively curated cDNA information, H-DBAS provides a unique platform for the analysis of complex AS.

INTRODUCTION

Alternative splicing (AS) is a phenomenon in which a single gene produces various functional protein isoforms. AS is frequently observed especially in higher eukaryotes. At least 50% of human genes are reported to be subjected to AS. However, the biological significance of this high level of AS and its regulation mostly remain elusive (1,2). For better understanding of AS in humans, we constructed a human-transcriptome database for alternative splicing (H-DBAS) in 2006, which collects information of human AS variants from the viewpoints of protein functions affected by AS. H-DBAS is based on the manually inspected and well-annotated cDNA information collected by the H-Invitational cDNA Annotation Project. By utilizing the annotation information and cDNA sequence information, it was possible to identify AS events that invoke changes in protein-coding regions, thereby influencing protein functions (3–5). Based on the result of intensive annotations of AS events, H-DBAS presents thousands of AS events that may increase the functional diversification of the human genome.

However, we further examined the evolutionary conservation of the identified AS events and found that a large number of these annotated AS events may not be evolutionarily conserved between humans and mice. Similar results were also reported by other groups (6). Our concern was that they could simply represent intrinsic noise of transcription inherently occurring in the human genome without biological relevance. Therefore, further extensive annotations in which AS events are likely to be translated into proteins and whether such AS events are evolutionarily conserved would be essential. Such information will be extremely useful to prioritize targets for future functional characterization of AS events and to determine the direction of validation experiments.

The latest generation of sequencers have greatly improved the cost and speed of cDNA sequencing (7). A recent paper reported the use of a new generation sequencer for in-depth identification and characterization of human AS events. They generated dozens of millions of shotgun RNA sequence tags by the so-called RNA-Seq analysis and analyzed the collected tags (RNA-Seq tags) to detect positions and frequencies of the usage of every splice junction (8,9). In this particular study, polyA+ RNA was used for RNA-Seq analysis. However, several methodological improvements have been made so that it is now possible to consider a similar approach for analysis of RNAs from any population. In a very recent study, we generated a total of 150-million RNA-Seq tag sequences

*To whom correspondence should be addressed. Tel: +81 3 3599 8800; Fax: +81 3 3599 8801; Email: t.imanishi@aist.go.jp using RNAs that were separately extracted from cytoplasmic, nuclear and polysome (translating ribosome) subcellular fractions in DLD-1 cells, a colon cancer cell line. In this update of H-DBAS, we incorporated this RNA-Seq data enabling a clear representation of which RNAs and their AS variants are identified in which subcellular fractions. Observing a particular AS variant in the polysome fraction should be especially important because it provides direct evidence for its translational consequence. Also, to determine whether an AS is evolutionarily conserved, we used a comparative genomic viewer. In this viewer, AS events are categorized according to whether they are transcribed from conserved genomic regions or whether the corresponding transcripts that are also identified in mice. The updated H-DBAS including these two expanded features should provide a unique and important resource to explore the complex world of human AS.

NEW FEATURES

Statistics of the new RNA-Seq datasets

By RNA-Seq analysis using Illumina GA, we generated 46 354 139, 47 120 831 and 54 901 628 single-end-read 36-bp RNA-Seq tags from cytoplasmic, nuclear and polysome fractions of the RNAs from DLD-1 cells, respectively. Separation of the respective subcellular fractions was confirmed by western blotting of glyceraldehyde-3-phosphate dehydrogenase, a cytoplasmic protein and lamin A/C, a nuclear protein, as well as real-time RT-PCR analysis of sno/scaRNAs, nuclear RNAs (see RNA-Seq analysis page on the top page of H-DBAS for the related experimental data; details of the experimental procedures are also described there). The RNA-Seq tags obtained were mapped to the reference human genome of UCSC genome browser (hg18) (10). To identify tags that span splice junctions, we used Eland RNA and TopHat (version 1.0.9) (11,12) with the default options of considering only junctions following the ‘GT–AG’ rule and allowing up to two base mismatches. We further selected the splice junctions that were supported by two or more RNA-Seq tags. As a result, 201 280, 236 764 and 319 577 junctions were represented in the RNA-Seq datasets derived from cytoplasmic, nuclear and polysome subcellular fractions, respectively (Table 1).

The RNA-Seq tag information obtained was further correlated with transcript information. For analyzing the subdataset of human AS variants, we used RefSeq transcripts (release 23) (13). Among the total of 26 814 human RefSeq transcripts, 10 923 were annotated to represent mutual AS variants according to H-InvDB (release 6.0) [see ref. (4) for further details]. In total, 81 547, 85 923 and 90 900 exons were represented by RNA-Seq tags derived from cytoplasmic, nuclear and polysome fractions, respectively. In addition, 47 615, 47 260 and 51 041 splice junctions were represented in the RNA-Seq tags in the respective fractions. Of these, 1067, 1021 and 1114 junctions corresponded to mutual AS junctions, directly suggesting that these AS events are expressed and located in the respective subcellular locations. Statistical analysis of the enrichment of tags also showed that some AS variants were enriched in a given subcellular location: 260, 254 and 299 AS variants were selectively observed in the cytoplasmic, nuclear and polysome fractions, respectively. Especially for 178 AS variant pairs, both of the variants appeared to be translated to proteins simultaneously in DLD-1 cells. All of the above extensive annotations on the biological relevance of each AS are represented as a graphic interface as described below.

RNA-Seq viewer

RNA-Seq viewer can be accessed from the RNA-Seq analysis page at the H-DBAS top page. On the RNA-Seq analysis page, RNA-Seq and AS annotation information were described in a table. In the table, the number of corresponding RNA-Seq tags and presumed subcellular locations of AS events were shown. By following the link from the table, details of the RNA-Seq tag supports in the junction appear in the RNA-Seq viewer. In this viewer, RefSeq transcripts and tags located in the splice junctions are represented. RNA-Seq tag information was further categorized so that users can examine tag distribution in each subcellular location. Figure 1 exemplifies RNA-Seq tag analysis in the case of caspase 4, an apoptosis-related cysteine peptidase gene. In this gene, the AS junction (indicated by a red line) was exclusively identified in nuclear fractions. Figure 1 also represents 35 RNA-Seq tags mapped to the corresponding splice junctions. These results suggested that the AS variant using the most upstream exon (using splice junctions marked in red) is retained in the nucleus and is not used for protein translation in DLD-1 cells.

Comparative genomics viewer

In order to distinguish AS events having a clear biological significance, it would be informative to consider whether an AS is evolutionarily conserved, for which we newly

Table 1. Statistics of human RefSeq junctions expressed in each cellular fraction using RNA-Seq

	Total RNA-Seq tags	RNA-Seq tags mapped to RefSeq regions	Represented exons	Represented splice junctions	Represented AS junctions
Cytoplasm	46 354 139	28 906 833	81 547	47 615	1067
Nuclear	47 120 831	28 939 028	85 923	47 260	1021
Polysome	54 901 628	29 720 537	90 900	51 041	1114

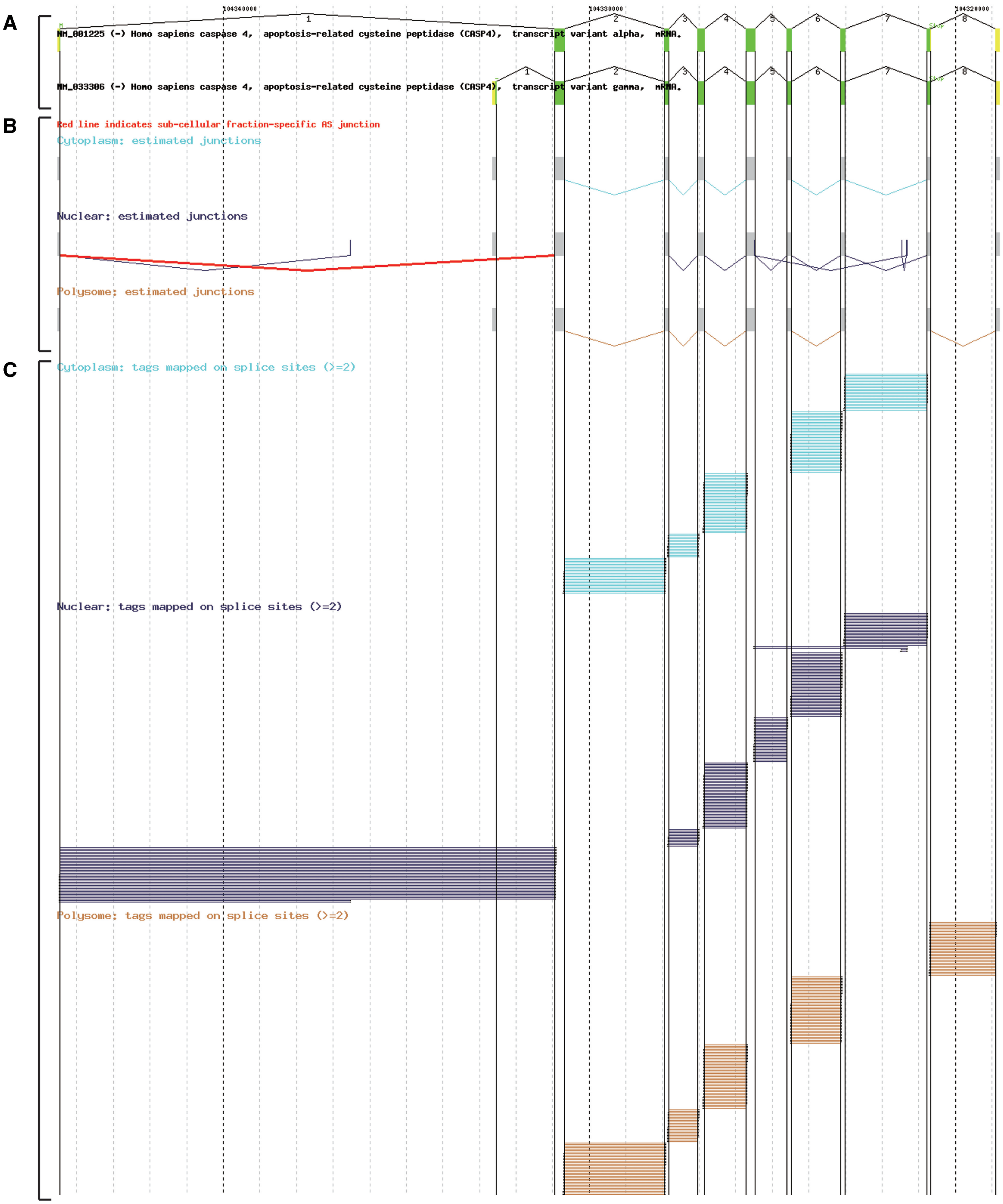

A
104340000
104330000
104320000
1
2
3
4
5
6
7
8
NM_001225 (-) Homo sapiens caspase 4, apoptosis-related cysteine peptidase (CASP4), transcript variant alpha, mRNA.
NM_033306 (-) Homo sapiens caspase 4, apoptosis-related cysteine peptidase (CASP4), transcript variant gamma, mRNA.
B
Red line indicates sub-cellular fraction-specific AS junction
Cytoplasm: estimated junctions
Nuclear: estimated junctions
Polysome: estimated junctions
C
Cytoplasm: tags mapped on splice sites (>=2)
Nuclear: tags mapped on splice sites (>=2)
Polysome: tags mapped on splice sites (>=2)

Figure 1. Screenshot of RNA-Seq viewer. Genomic regions in caspase 4, an apoptosis-related cysteine peptidase (CASP4) and the estimated junctions with the two or more supporting RNA-Seq tags mapped to the corresponding genomic regions. (**A**) AS variants of RefSeq are represented. Annotated protein-coding regions and untranslated regions are indicated by green and yellow boxes, respectively. (**B**) Junctions estimated by the mapped RNA-Seq tags derived from cytoplasmic, nuclear and polysome cellular fractions are shown in cyan, navy and brown, respectively. If the AS junction of RefSeq transcript is expressed in unique sub-cellular fraction (nuclear in this figure), it is shown in red. The gray boxes indicate the assembled exonic regions of RefSeq transcripts. (**C**) RNA-Seq tags which support the junction are represented. Two or more RNA-Seq tags mapped on the splice sites are shown by each sub-cellular fraction. The represented colors are the same as (B).

implemented a comparative genomics viewer to empirically represent the degree of evolutionary conservation for any AS. In this viewer, each AS variant can be viewed for the following points: (i) whether its surrounding genomic sequence is conserved between humans and mice and (ii) whether the corresponding AS event is also observed in mice. Genomic sequences and alignment information were obtained from UCSC genome browser (hg18 and mm9 for humans and mice, respectively) (10). For full-length cDNA information, we used 65 158 human full-length cDNAs and 122 544 mouse full-length cDNAs from H-InvDB (5), FANTOM (14) and Mammalian Gene Collection (15). In total, 20 803 representative AS variants (RASVs) among all human full-length cDNAs are represented. Among 207 399 exons of the total 20 803 human RASVs, 27 567 exons were mapped to the genomic regions that had no aligned mouse genomic regions. On the other hand, 22 396 exons were mapped to the aligned genomic regions (coverage ≥70% and identity ≥60%), but the corresponding transcripts were not identified in mice. The remaining 157 436 exons were mapped to the conserved genomic regions and corresponding transcripts were identified in mouse full-length cDNAs. Among the 7875 conserved RASVs thus identified, 5494 were equally spliced variants (ESVs) with mouse full-length cDNAs, which are conserved between humans and mice and are likely to have evolutionarily conserved biological roles (Table 2). For example, as shown in Figure 2, the phosphoinositide-3-kinase regulatory subunit gene has several AS variants. For the two AS variants, their splice patterns are identical to those of the mouse full-length cDNAs. These AS variants may contribute to functional diversification of gene function, playing conserved biological roles both in humans and mice. Further details of the statistical analysis of frequencies of conserved AS variants in various gene groups have been described previously (16). The comparative genomics viewer is embedded in the main AS viewer. It can also be accessed from the summary annotation table at the H-DBAS top page and users can search specifically about the comparative genomics analysis from Advanced search page at the top page.

Table 2. Statistics of comparative genomics between human and mouse full-length cDNAs

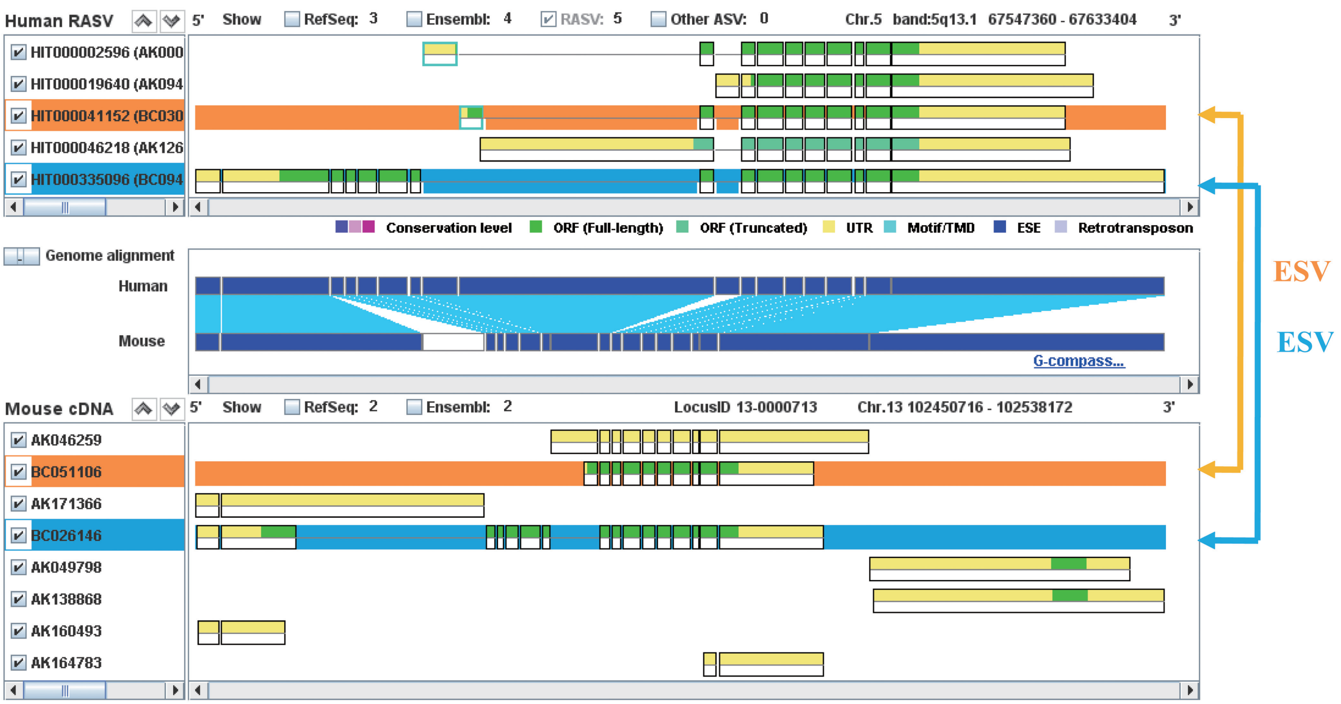
	At least one exon conserved	ESV	Conserved AS
RefSeq	10 217	4193	392
RASV	7875	5494	499

RASV: representative AS variant; ESV: equally spliced variant.

FUTURE PERSPECTIVES

We updated our H-DBAS so that AS transcripts having various types of annotation information can be represented in an integrative manner. These types of

phosphatidylinositol 3-kinase regulatory subunit alpha (PI3-kinase p85 subunit alpha) gene



Human RASV 5' Show ☐ RefSeq: 3 ☐ Ensembl: 4 ☑ RASV: 5 ☐ Other ASV: 0 Chr.5 band:5q13.1 67547360 - 67633404 3'

☑ HIT000002596 (AK000
☑ HIT000019640 (AK094
☑ HIT000041152 (BC030
☑ HIT000046218 (AK126
☑ HIT000335096 (BC094

Mouse cDNA 5' Show ☐ RefSeq: 2 ☐ Ensembl: 2 LocusID 13-0000713 Chr.13 102450716 - 102538172 3'

☑ AK046259
☑ BC051106
☑ AK171366
☑ BC026146
☑ AK049798
☑ AK138868
☑ AK160493
☑ AK164783

Figure 2. Screenshot of comparative genomic viewer. AS variants in the phosphatidylinositol 3-kinase regulatory subunit alpha (PI3-kinase p85 subunit alpha) gene are shown both in humans and mice. The exon structures of the human AS variants and the mouse full-length cDNAs are shown in the upper and lower panels, respectively, across the human–mouse genome alignment. In this view option (Exon view), constitutively spliced introns of the transcripts are omitted. Mutually equally spliced variants in humans and mice are indicated by blue and orange arrows, respectively.

information include manual annotations; full-length cDNA sequences; RNA-Seq tags derived from RNAs extracted from nuclear, cytoplasm and polysome fractions; and degree of evolutionary conservation of AS. By enabling the integrative interpretation of annotation information, we believe that H-DBAS can serve as a unique and useful database for future functional characterization of AS events. In future, we aim to further enrich the diverse annotation information connected to each AS. For this purpose, we aim to expand similar RNA-Seq analysis to cover the transcriptome information of mice and other mammals. Also, we aim to continue to collect RNA-Seq tags from a wider variety of cell types cultured under different conditions in order to understand which AS events are transcribed in which cell types and under what cellular conditions. Results of such extensive analyses will be fed back to the manual annotations in H-InvDB. With integrative transcriptome data, we aim to provide expanded knowledge of the biological significance of the functional diversification of human genes realized by AS, which should add useful molecular background to the complex human gene network created by a limited number of genes.

ACKNOWLEDGEMENTS

The authors thank Y. Kawahara, A. Matsuya, H. Nakaoka, T. Habara, F. Todokoro and C. Yamasaki for their assistance in genome mapping and ORF prediction. We also thank E. Sekimori for the technical support for RNA-Seq analysis, M. Nitta for constructing the comparative genomics viewer, and T. Endo for the technical support for server usage. Finally, we are grateful to all those who annotated the full-length human cDNAs at the H-Invitational and H-Invitational two conferences.

FUNDING

Integrated database project of the Ministry of Economy, Trade, and Industry of Japan, Ministry of Education, Culture, Sports, Science and Technology of Japan, National Institute of Advanced Industrial Science and Technology (AIST) and Japan Biological Informatics Consortium (JBIC). Funding for open access charge: AIST.

Conflict of interest statement. None declared.

REFERENCES

1. Modrek,B. and Lee,C. (2002) A genomic view of alternative splicing. *Nat. Genet.*, **30**, 13–19.
2. Tress,M.L., Martelli,P.L., Frankish,A., Reeves,G.A., Wesselink,J.J., Yeats,C., Olason,P.L., Albrecht,M., Hegyi,H., Giorgetti,A. *et al.* (2007) The implications of alternative splicing in the ENCODE protein complement. *Proc. Natl Acad. Sci. USA*, **104**, 5495–5500.
3. Imanishi,T., Itoh,T., Suzuki,Y., O'Donovan,C., Fukuchi,S., Koyanagi,K.O., Barrero,R.A., Tamura,T., Yamaguchi-Kabata,Y., Tanino,M. *et al.* (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.*, **2**, e162.
4. Takeda,J., Suzuki,Y., Nakao,M., Kuroda,T., Sugano,S., Gojobori,T. and Imanishi,T. (2007) H-DBAS: alternative splicing database of completely sequenced and manually annotated full-length cDNAs based on H-Invitational. *Nucleic Acids Res.*, **35**, D104–D109.
5. Yamasaki,C., Murakami,K., Fujii,Y., Sato,Y., Harada,E., Takeda,J., Taniya,T., Sakate,R., Kikugawa,S., Shimada,M. *et al.* (2008) The H-Invitational database (H-InvDB), a comprehensive annotation resource for human genes and transcripts. *Nucleic Acids Res.*, **36**, D793–D799.
6. Modrek,B. and Lee,C.J. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat. Genet.*, **34**, 177–180.
7. Graveley,B.R. (2008) Molecular biology: power sequencing. *Nature*, **453**, 1197–1198.
8. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
9. Pan,Q., Shai,O., Lee,L.J., Frey,B.J. and Blencowe,B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, **40**, 1413–1415.
10. Karolchik,D., Kuhn,R.M., Baertsch,R., Barber,G.P., Clawson,H., Diekhans,M., Giardine,B., Harte,R.A., Hinrichs,A.S., Hsu,F. *et al.* (2008) The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res.*, **36**, D773–D779.
11. Trapnell,C., Pachter,L. and Salzberg,S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
12. 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.
13. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
14. Carninci,P., Kasukawa,T., Katayama,S., Gough,J., Frith,M.C., Maeda,N., Oyama,R., Ravasi,T., Lenhard,B., Wells,C. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
15. Gerhard,D.S., Wagner,L., Feingold,E.A., Shenmen,C.M., Grouse,L.H., Schuler,G., Klein,S.L., Old,S., Rasooly,R., Good,P. *et al.* (2004) The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome Res.*, **14**, 2121–2127.
16. Takeda,J., Suzuki,Y., Sakate,R., Sato,Y., Seki,M., Irie,T., Takeuchi,N., Ueda,T., Nakao,M., Sugano,S. *et al.* (2008) Low conservation and species-specific evolution of alternative splicing in humans and mice: comparative genomics analysis using well-annotated full-length cDNAs. *Nucleic Acids Res.*, **36**, 6386–6395.