Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W681–W685***
doi:10.1093/nar/gki407

ASePCR: alternative splicing electronic RT–PCR in multiple tissues and organs

Namshin Kim[1], Dajeong Lim, Sanghyuk Lee[1] and Heebal Kim*

School of Agricultural Biotechnology, Seoul National University, Seoul 151-742, Korea and
[1]Division of Molecular Life Sciences, Ewha Womans University, Seoul 120-750, Korea

Received February 14, 2005; Revised and Accepted March 21, 2005

ABSTRACT

RT–PCR is one of the most powerful and direct methods to detect transcript variants due to alternative splicing (AS) that increase transcript diversity significantly in vertebrates. ASePCR is an efficient web-based application that emulates RT–PCR in various tissues. It estimates the amplicon size for a given primer pair based on the transcript models identified by the reverse e-PCR program of the NCBI. The tissue specificity of each PCR band is deduced from the tissue information of expressed sequence tag (EST) sequences compatible with each transcript structure. The output page shows PCR bands like a gel electrophoresis in various tissues. Each band in the output picture represents a putative isoform that could happen in a tissue-specific manner. It also shows the EST alignment and tissue information in the genome browser. Furthermore, the user can compare the AS patterns of orthologous genes in other species. The ASePCR, available at http://genome.ewha.ac.kr/ASePCR/, supports the transcriptome models of the RefSeq, Ensembl, ECgene and AceView for human, mouse, rat and chicken genomes. It will be a valuable web resource to explore the transcriptome diversity associated with different tissues and organs in multiple species.

INTRODUCTION

Recent studies have indicated that alternative splicing (AS) is the most significant component of functional complexity in the human genome because about 40–60% of all human genes (1–4) and 74% of multi-exon human genes (5) have isoforms due to AS events. Identifying AS forms using an *in silico* approach depends on comparing large amount of ESTs with another set of sequences. The methods for identifying AS forms can be classified into those that compare ESTs with their representative transcriptional sequence, such as mRNA sequence (1), or align ESTs to their genomic sequences (2–4,6). Recently, the focus of AS research has been on dissecting the functional implications of AS in tissue-specificity (7,8), biological processes (9) and tumor-development (10), or on patterns of conservation among species (11). These studies have tried to interpret the specific biological impact of each AS variant. Due to the biological significance in these diverse subjects, many groups are trying to build reference databases on AS, such as the ASD (12), ASAP (13) and ECgene (14,15).

The RT–PCR is one of the most powerful methods for the detection of AS in various tissues. The RT–PCR is easier and more popular than microarray technique in terms of confirming AS variants of an individual gene. Generally, a targeted sequence is analyzed and whether there are AS events or not is confirmed in a laboratory on the basis of product sizes of RT–PCR in various tissues.

There have been several attempts to emulate the PCR experiments *in silico*. NCBI's e-PCR is the most well-known application that supports searches between sequence database and the sequence tagged site (STS) database (16). However, the search scope in the web version is mostly for the STS markers, not for general primer pairs. The UCSC genome center recently released an application called the 'UCSC *In-Silico* PCR' that searched the genomic DNA database with a pair of PCR primers (http://genome.ucsc.edu/cgi-bin/hgPcr/). No PCR-based program is available to explore the transcriptome diversity due to AS events.

In this manuscript, we present a novel web-based application, ASePCR, which emulates the RT–PCR experiment in various tissues. It uses the reverse e-PCR program to find the transcripts for a given primer pair and the gene expression pattern in multiple tissues is deduced from the tissue information of EST sequences compatible with the transcript structure. The result should provide a valuable insight on isoform expression pattern in multiple tissues and organs.

*To whom correspondence should be addressed. Tel: +82 28804803; Fax: +82 28732271; Email: heebal@snu.ac.kr

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors

MATERIALS AND METHODS

Datasets and programs

The reverse e-PCR program used to identify transcripts with given primer sequences was downloaded from the NCBI website (http://www.ncbi.nlm.nih.gov/sutils/e-pcr/, e-PCR 2.3.1) (16). The transcript models supported include the ECgene (http://genome.ewha.ac.kr/ECgene/) (14), AceView (http://www.ncbi.nih.gov/IEB/Research/Acembly/), Ensembl (http://www.ensembl.org/) (17), RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/) (18) and mRNA sequences in the GenBank. Models for the most recent genome builds were used, i.e. NCBI Build 35/hg17 for human, NCBI Build 33/mm5 for mouse, Baylor HGSC v3.1/rn3 for rat and CGSC Feb 2004/galGal2 for chicken.

To speed up the calculation enough to be suitable for a webserver application, all transcriptome models and EST sequences were pre-aligned against the genome using the BLAT program (19). The resulting alignments were stored in a database so that a simple database query could retrieve EST sequences overlapping with the amplified fragments in the RT–PCR experiment.

In an effort to increase the tissue coverage, we manually classified all cDNA libraries in the dbEST for human, mouse, rat and chicken. Tissue information was obtained from various resources, such as the Cancer Genome Anatomy Project (CGAP) (http://cgap.nci.nih.gov/), the UniGene library browser (http://www.ncbi.nih.gov/UniGene/) and the MEROPS database (http://merops.sanger.ac.uk/).

Orthologous region was identified on the basis of the Chain/Net track in the UCSC genome browser. The primer sequences within the orthologous sequences were identified from the sequence alignment obtained by the 'needle' program in the EMBOSS package (http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Apps/) (20). The default criterion to be a satisfactory primer in orthologous transcripts is the minimum percent identity of 60%.

Overview of the algorithm

The algorithm consists of four major steps as described below.

(i) Finding the transcript models in the target databases. We use the reverse e-PCR program of the NCBI to find the transcript models with the given primer pair sequences. Option of permitting 1 bp mismatch and/or 1 bp indel is supported to take imperfect priming into consideration. The program is fast enough to allow a web service since it is based on hash indexing of transcriptome sequences. The transcriptome databases supported are the RefSeq, Ensembl, ECgene, AceView and GenBank mRNA sequences.

(ii) Finding EST sequences that overlap with transcript models. Even though the BLAST search against the dbEST database would be the first choice, it is not fast enough to support the webserver application. Instead, we use the genomic alignment of transcript and EST sequences. In building the ECgene models that use genome-based EST clustering, all mRNA and EST sequences are aligned against the genome using the BLAT program. Only the best alignments are curated and stored in a database adopting the binning scheme (21). Genomic alignments of other transcriptome models are downloaded from the UCSC genome center. EST sequences aligned inside the genomic region defined by the primer pair sequences are retrieved by a simple database query. For each transcript model, compatible EST sequences, i.e. with no intron-exon mismatches, are found by examining the genomic alignments.

(iii) Tissue information and PCR band assignment. For each transcript model, compatible EST sequences found in the previous step are used to determine the expression pattern in multiple tissues and organs. Specifically, we calculate the 'EST coverage' for each transcript in each tissue. The definition of EST coverage is the ratio of the number of nucleotides covered by ESTs and the number of nucleotides of the PCR product generated from a given primer pair. If the value exceeds the cutoff number specified by the user, it is assumed that the transcript model is expressed in the tissue of interest and a PCR band at the amplicon size is drawn for the corresponding tissue. Tissue EST coverage of 100% means that the whole amplicon region has the EST evidence from the tissue of interest. Tissue EST coverage of 1% would show a PCR band for tissues with a minimal number of ESTs (usually 1 EST). The EST coverage is a measure of likelihood of real expression of a given form of transcript but is not a reflection of the expression level. Rather it is the number of EST clones that would more accurately reflect the expression level of each transcript.

(iv) Ortholog search by pair-wise alignment. The ortholog search feature allows the user to examine the AS pattern in orthologous genes of various species. To identify the orthologous transcripts in other organisms, we used the comparative genomics track 'Chain/Net' in the UCSC genome browser. For each transcript model found in step 1, orthologous transcripts are identified using the conserved genomic blocks between two organisms. The primer sequences are not always conserved perfectly in other species. So we use the Needleman–Wunsh algorithm to align two primer sequences against the orthologous transcripts. Steps 2 and 3 are repeated for other organisms with the orthologous transcripts and the new primer sequences are identified.

RESULTS

Input format

Figure 1 shows the GUI for the ASePCR input page. Most items are self-explanatory, but a user's guide is available on the website for more details. Four model organisms are currently being supported and the tissue classification can be displayed for each organism. The user should select the particular organism and supply primer information—the forward and reverse primers. Flipping the reverse primer is supported for user convenience. Primer sequences must be longer than 15 bp. The primer quality option supports 1 bp mismatch and/or 1 bp indel to compensate for imperfect primers. Users can also specify the product size range.

The minimum EST coverage is a critical parameter. The whole amplicon region may not have full EST coverage for the tissue of interest due to limited EST data. ASePCR draws PCR bands for transcripts whose EST coverage in each tissue is larger than the cutoff value. Therefore, a small cutoff value allows more bands to appear in the final output at the risk of producing false positives. Selecting the tissue EST coverage of 100% shows PCR bands only for transcript models whose

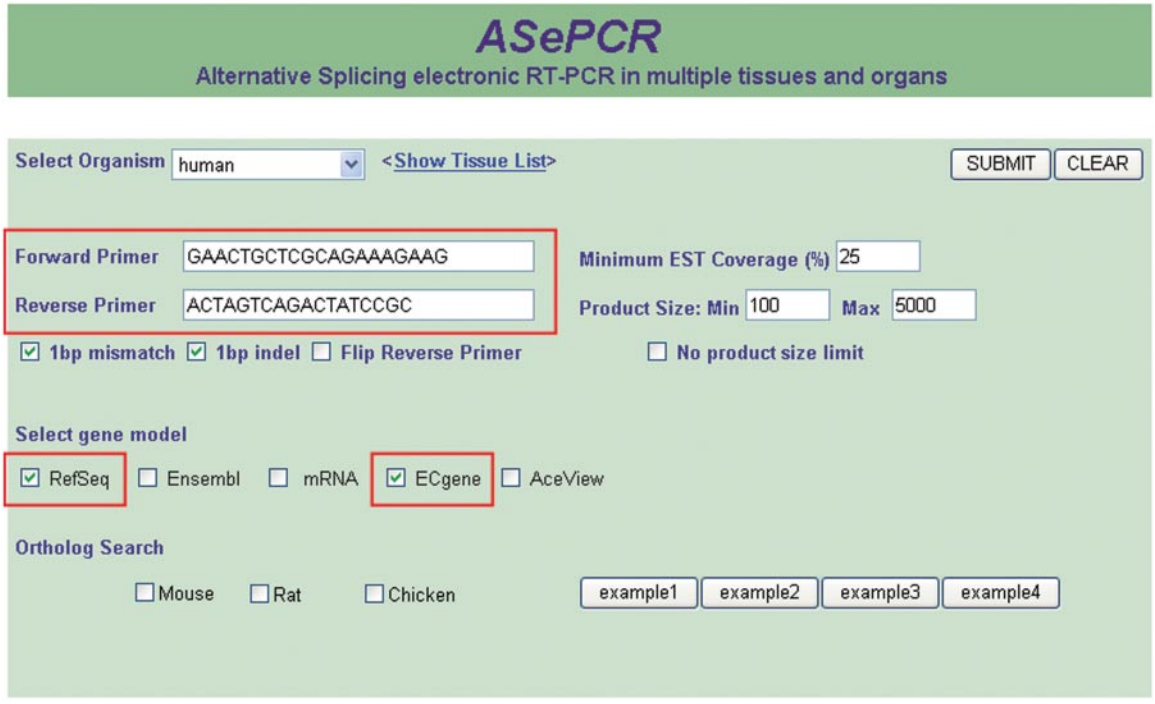

Figure 1. Input GUI design for ASePCR. Essential fields are the organism, primer sequences, minimum EST coverage and gene models.

amplicon can be reconstructed by assembling EST sequences from the corresponding tissue.

The user then selects the transcriptome models. The RefSeq, Ensembl and GenBank mRNA sequences are the most widely used databases of reference genes. However, they do not reflect the transcript variation due to AS properly. Gene models in the ECgene and AceView are built from the genomic alignment of mRNA and ESTs, thereby providing more extensive analysis of AS. These gene models should be selected to find AS variants with the EST evidence only. AceView annotation is available only for the human genome. The ortholog search option is useful to compare the AS pattern in other organisms in case the primer sequences are conserved with percent identity over 60%.

Even though choosing many gene models and species for ortholog search is certainly more informative, it requires substantially more calculation time. It is recommended to use only the models and organisms of interest.

Output features

The output page consists of several sections that include the input summary, transcript search result, RT–PCR image, genome browser showing EST alignments, EST information and the transcript sequences in the FASTA format. Figure 2A shows the RT–PCR image for the BTF3 (basic transcription factor 3) gene as an example of the *in-silico* RT–PCR. It represents the simulated gels for various tissues. PCR bands for the transcripts with the primer sequences are drawn whenever the tissue EST coverage is larger than the cutoff value. The first gel, labeled as 'All Tissues', shows all bands (amplicons) regardless of the tissue origin and EST coverage. Gels with a specific tissue name show transcripts that are presumably expressed in that tissue, judging by the tissue EST coverage. The last gel, labeled as 'others', implies expression in unclassified tissues. The numbers of cDNA libraries and ESTs are indicated at the bottom of the gel. The darkness of PCR bands indicates the relative expression level of transcripts. Clicking on each band opens a new window that provides more detailed information on the transcript and EST sequences. We also provide the option of customizing the gel image as shown in Figure 2A. Users can adjust the band range and the minimum EST coverage after examining the EST coverage for all bands.

The EST alignment in the genome browser, shown in Figure 2B, is particularly useful. Drawing the PCR band according to the minimum coverage can be misleading sometimes. In case the primer sequences are not carefully designed, transcript may have high EST coverage due to ESTs that are not related to the specific AS event. This problem can be remedied by examining the EST alignment. We developed our own genome browser that is quite similar to the UCSC genome browser. Note that the tissue source of EST sequences is specified and that the EST sequences are colored according to the tissue origin. The primer positions are specified in red vertical lines or boxes (if the primer alignment spans multiple exons).

Detailed information on EST sequences is given in the EST information part. The tissue information is summarized with the links to the cDNA library description. We also show the multiple alignment of transcripts so that the user may recognize the difference easily. Transcript sequences are given in the last part of the FASTA format with the primer sequences specified in color. If the user selects the ortholog search, ASePCR repeats the calculation for the orthologous transcripts and shows the result in a similar fashion.

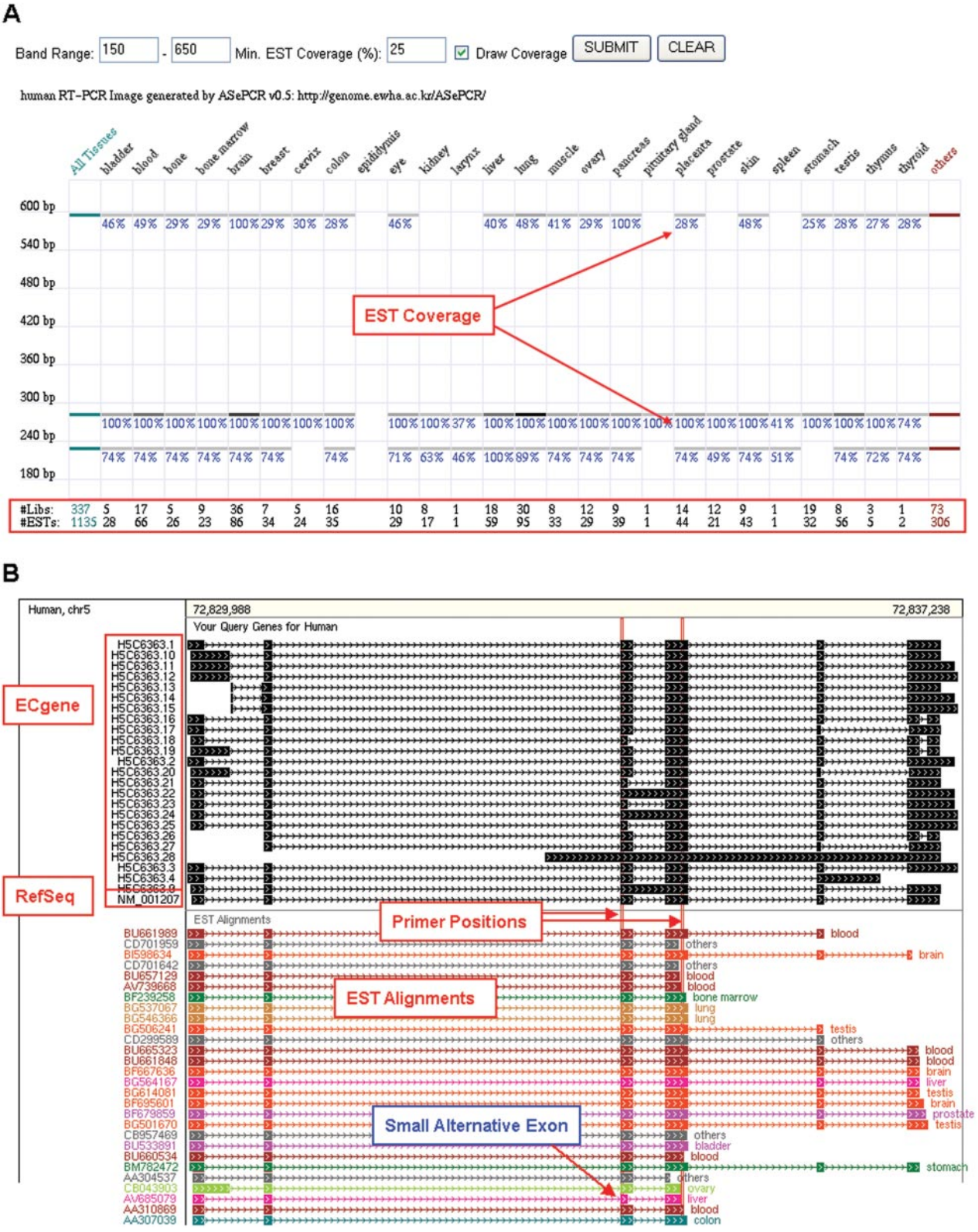

Figure 2. Part of the output page from the ASePCR. Primer sequences are designed to show the AS pattern of BTF3 (basic transcription factor 3) gene. Selected gene models are the RefSeq and ECgene. (**A**) The simulated RT–PCR image shows the expression pattern of AS variants in multiple tissues. The option of showing EST coverage is selected in this customized picture. Darkness of each PCR band is proportional to the number of EST sequences indicated in the lower box. Note that BTF3 gene has 1135 ESTs overlapping inside the primer region. (**B**) Transcript models, primer positions, EST alignments and their tissue source are shown in the custom-made genome browser that resembles the UCSC genome browser. Two exons appear within the transcript region defined by the primer pair. Whereas RefSeq has only one transcript model (281 bp), ECgene shows two additional patterns of AS, i.e. transcripts with the shorter first exon (227 bp) and with the intron retention (595 bp). An EST from liver (AV685079) is compatible with the transcript generating the PCR band at 227 bp.

DISCUSSION

In this paper, we presented an efficient web application that emulates the RT–PCR experiment to identify AS variants. Simple interface and fast program speed would allow users to explore the transcript variants in multiple tissues efficiently.

However, it should be emphasized that the PCR bands are assigned for each tissue according to the EST evidence only. EST-based predictions may be different from the experimental results due to several reasons (22) and the user should bear in mind that the prediction can be wrong especially if the transcripts do not have sufficient EST sequence data. For example, the number of EST sequences may not reflect the expression level properly since many public cDNA libraries are prepared by normalized or subtracted protocols. Furthermore, PCR bands with low EST coverage should be critically reviewed since the relevant ESTs may not reflect the splicing pattern of interest. In the example of BTF3 in Figure 2, the transcript with the small alternative exon giving a band at 227 bp is expressed in liver since its EST coverage is 100%. Scrolling down the EST alignment reveals that an EST from liver (AV685079) is responsible for the transcript. Other bands below the EST coverage of 74% are from EST sequences overlapping with the second exon only and should not be regarded as a genuine proof for the small alternative exon. In order to concentrate on transcripts with full EST evidence, the user should use 100% as the cutoff value of EST coverage. The transcript responsible for the band at 227 bp is expected to be liver-specific in this standard. Similarly, transcript with intron retention is expressed specifically in brain and pancreas. Therefore, the EST coverage is not an inherently conclusive statistical measure but would be a valuable guide that users may take into consideration in evaluating specific cases.

Even though ASePCR was developed to visualize the RT–PCR result in search of transcript variants, it should be helpful for bench biologists to test the primer specificity since it searches the entire transcriptome models. In the near future, we plan to add features to show the variation of AS patterns according to different pathological states, such as cancer and normal tissues.

ACKNOWLEDGEMENTS

We thank Dr Robert Klein for helpful comments and English editing on the manuscript. This work was supported by the Bioinformatics Research Program of the Ministry of Science and Technology of Korea (to S.L., grant 2005-00201) and the BioGreen 21 Program of the Korean Rural Development Administration (to H.K.) and the Brain Korea 21 Project of the Ministry of Education (to H.K.). Funding to pay the Open Access publication charges for this article was provided by BioGreen 21 Program.

Conflict of interest statement. None declared.

REFERENCES

1. Brett,D., Hanke,J., Lehmann,G., Haase,S., Delbruck,S., Krueger,S., Reich,J. and Bork,P. (2000) EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett*., **474**, 83–86.
2. Croft,L., Schandorff,S., Clark,F., Burrage,K., Arctander,P. and Mattick,J.S. (2000) ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nature Genet*., **24**, 340–341.
3. Mironov,A.A., Fickett,J.W. and Gelfand,M.S. (1999) Frequent alternative splicing of human genes. *Genome Res*., **9**, 1288–1293.
4. Modrek,B. and Lee,C. (2002) A genomic view of alternative splicing. *Nature Genet*., **30**, 13–19.
5. Johnson,J.M., Castle,J., Garrett-Engele,P., Kan,Z., Loerch,P.M., Armour,C.D., Santos,R., Schadt,E.E., Stoughton,R. and Shoemaker,D.D. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **302**, 2141–2144.
6. Kim,N., Shin,S. and Lee,S. (2004) ASmodeler: gene modeling of alternative splicing from genomic alignment of mRNA, EST and protein sequences. *Nucleic Acids Res*., **32**, W181–W186.
7. Xu,Q., Modrek,B. and Lee,C. (2002) Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res*., **30**, 3754–3766.
8. Yeo,G., Holste,D., Kreiman,G. and Burge,C.B. (2004) Variation in alternative splicing across human tissues. *Genome Biol*., **5**, R74.
9. Kan,Z., States,D. and Gish,W. (2002) Selecting for functional alternative splices in ESTs. *Genome Res*., **12**, 1837–1845.
10. Xu,Q. and Lee,C. (2003) Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. *Nucleic Acids Res*., **31**, 5635–5643.
11. Thanaraj,T.A., Clark,F. and Muilu,J. (2003) Conservation of human alternative splice events in mouse. *Nucleic Acids Res*., **31**, 2544–2552.
12. Thanaraj,T.A., Stamm,S., Clark,F., Riethoven,J.J., Le Texier,V. and Muilu,J. (2004) ASD: the alternative splicing database. *Nucleic Acids Res*., **32**, D64–D69.
13. Lee,C., Atanelov,L., Modrek,B. and Xing,Y. (2003) ASAP: the alternative splicing annotation project. *Nucleic Acids Res*., **31**, 101–105.
14. Kim,P., Kim,N., Lee,Y., Kim,B., Shin,Y. and Lee,S. (2005) ECgene: genome annotation for alternative splicing. *Nucleic Acids Res*., **33**, D75–D79.
15. Kim,N., Shin,S. and Lee,S. (2005) ECgene: genome-based EST clustering and gene modeling for alternative splicing. *Genome Res*., **15**, 566–576.
16. Rotmistrovsky,K., Jang,W. and Schuler,G.D. (2004) A web server for performing electronic PCR. *Nucleic Acids Res*., **32**, W108–W112.
17. Hubbard,T., Andrews,D., Caccamo,M., Cameron,G., Chen,Y., Clamp,M., Clarke,L., Coates,G., Cox,T., Cunningham,F. *et al*. (2005) Ensembl 2005. *Nucleic Acids Res*., **33**, D447–D453.
18. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*., **33**, D501–D504.
19. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res*., **12**, 656–664.
20. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*., **16**, 276–277.
21. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al*. (2003) The UCSC Genome Browser Database. *Nucleic Acids Res*., **31**, 51–54.
22. Gupta,S., Zink,D., Korn,B., Vingron,M. and Haas,S.A. (2004) Strengths and weaknesses of EST-based prediction of tissue-specific alternative splicing. *BMC Genomics*, **5**, 72.