Published online 26 August 2010

Nucleic Acids Research, 2010, Vol. 38, No. 20 ***e187***
doi:10.1093/nar/gkq754

Verifying expressed transcript variants by detecting and assembling stretches of consecutive exons

Tzu-Hung Hsiao[1,2], Chien-Hong Lin[2], Te-Tsui Lee[2], Ji-Yen Cheng[3], Pei-Kuen Wei[3], Eric Y. Chuang[1,*] and Konan Peck[2,*]

[1]Departmant of Electrical Engineering, National Taiwan University, Taipei, Taiwan 106, [2]Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan 115 and [3]Research Center for Applied Sciences, Academia Sinica, Taipei, Taiwan 115, ROC

Received February 9, 2010; Revised August 4, 2010; Accepted August 6, 2010

ABSTRACT

We herein describe an integrated system for the high-throughput analysis of splicing events and the identification of transcript variants. The system resolves individual splicing events and elucidates transcript variants via a pipeline that combines aspects such as bioinformatic analysis, high-throughput transcript variant amplification, and high-resolution capillary electrophoresis. For the 14 369 human genes known to have transcript variants, minimal primer sets were designed to amplify all transcript variants and examine all splicing events; these have been archived in the ASprimerDB database, which is newly described herein. A high-throughput thermocycler, dubbed GenTank, was developed to simultaneously perform thousands of PCR amplifications. Following the resolution of the various amplicons by capillary gel electrophoresis, two new computer programs, AmpliconViewer and VariantAssembler, may be used to analyze the splicing events, assemble the consecutive exons embodied by the PCR amplicons, and distinguish expressed versus putative transcript variants. This novel system not only facilitates the validation of putative transcript variants and the detection of novel transcript variants, it also semi-quantitatively measures the transcript variant expression levels of each gene. To demonstrate the system's capability, we used it to resolve transcript variants yielded by single and multiple splicing events, and to decipher the exon connectivity of long transcripts.

INTRODUCTION

The generation of transcript variants by alternative splicing is a common process in human gene expression; indeed, >70% of human genes are known to express transcript variants (1). This mechanism allows a large repertoire of proteins to be generated from a limited number of genes. Various studies have shown that the transcript variants of a gene may have opposing roles, and alternative splicing is known to be a key factor in cancer progression. For example, Bcl-x, which is associated with cell survival/apoptosis, has two isoforms, Bcl-xL and Bcl-xS. The longer Bcl-xL isoform acts as an apoptotic inhibitor, whereas the shorter form acts as an apoptotic activator (2).

Several approaches have been used to identify the putative transcript variants of a given gene. Most such approaches have been based on bioinformatic analysis of expressed sequence tag (EST) sequences, in which a gene's EST sequences are aligned with its reference genomic sequence and all possible splicing events are identified from there (3–5). Another, higher-throughput empirical strategy is microarray analysis utilizing exon- or exon-exon junction-specific probes to measure exon expression and identify splicing events; in this method, the expression levels of individual exons and the between-sample differential expression patterns of splicing events are measured (1,6,7).

Recently, 'next generation' genome-sequencing technologies have been used to uncover splicing events in different organs and tissues (8–10). These genome-sequencing technologies generate huge cDNA datasets consisting of tens of millions of short sequencing reads. These datasets provide comprehensive surveys of splicing complexities in different tissues and contain vast quantities of data on putative splice junctions

*To whom correspondence should be addressed. Tel: +886 2 2652 3072; Fax: +886 2 2785 8594; Email: konan@ibms.sinica.edu.tw
Correspondence may also be addressed to Eric Y. Chuang. Tel: +886 2 3366 3660; Fax: +886 2 3366 3682; Email: chuangey@ntu.edu.tw

(i.e. putative splicing events). However, the large number of putative splicing events that may be identified in this way poses a great challenge for researchers seeking to examine the expression of transcript variants.

For example, in order to infer the biological functions of a transcript variant, it is necessary to decipher the exon connectivity of a variant, i.e. the composition and order of exons of a variant, and assemble exons into full-length transcripts. Methods to construct sets of putative full-length transcript variants by permuting the splicing events identified in EST libraries have been reported (11). However, if a gene has two potential splicing events, then the permutation method yields four putative transcript variants, and so forth, often creating a huge number of variants that must be identified. In order to minimize the number of putative transcript variants, several statistical methods have been used to estimate the probability that each putative variant encodes a viable transcript (5,12). In another study, an algorithm was developed to predict the number of transcript variants and compute for the putative full-length transcripts based on the hybridization signals of an exon microarray (13). In other work, detailed analyses of the hybridization signals from individual exon-specific probes have been used to estimate the putative expression profiles for sets of transcript variants (14–16). The accuracy of this strategy is decreased, however, if the splicing events are characterized by complex connectivity or if novel transcript variants are present. False predictions can be eliminated by experimental approaches such as the use of RT–PCR followed by electrophoresis to check the sizes of the various amplicons. However, such studies employ PCR primer pairs designed to amplify each exon–exon junction of the gene under investigation, and this strategy was found to require an average of 33 RT–PCR reactions per gene (17).

Here, we describe an integrated system for high-throughput transcript variant analysis that is capable of resolving individual transcript variants by combining bioinformatic analysis, RT-PCR, and capillary electrophoresis. This system not only detects stretches of consecutive exons encompassing multiple splicing events, it also semi-quantitatively measures the expression of each transcript variant. In brief, we collected all the splicing events archived in various databases, and used this information to design minimal PCR primers sets for each gene, such that all the variously sized transcript variants could be resolved by a multi-channel capillary electrophoresis instrument capable of carrying out thousands of reactions per day. Capillary electrophoresis has been shown to be a precise and efficient method for detection of splicing events (18). Our integrated system provides a high-throughput solution for validating splicing events, exon connectivity, and putative transcript variants that have been identified with *in silico* approaches or empirical techniques [e.g. exon microarrays and next-generation sequencing (NGS)].

MATERIALS AND METHODS

Samples and PCR amplification

Total RNA was derived from a lung adenocarcinoma tissue sample purchased from Clontech (Mountain View, CA). Fluorescent FAM dye-labeled reverse primers were synthesized with an in-house constructed oligonucleotide synthesizer, as previously described (19). Total RNA (1 μg) was reverse transcribed at 50°C for 1 h using 0.5 pmol of oligo-dT primer and reverse transcriptase III (Invitrogen, Carlsbad, CA) in a total volume of 11 μl. PCR was performed using 1 μl of the reverse-transcribed (RT) cDNA and a PCR reaction mixture containing 0.5 μM of the paired primers, 0.24 mM dNTPs, 20 mM Tris–HCl, pH 8.4, 50 mM KCl, 1.4 mM $MgCl_2$, and 0.5 units of Platinum *Taq* DNA polymerase (Invitrogen) in a total volume of 25 μl. The PCR conditions were as follows: initial denaturation at 94°C for 2 min, followed by 35 cycles of 94°C for 40 s, 57°C for 40 s, and 72°C for 2 min, and a final extension at 72°C for 10 min. All primers used in this study are shown in Supplementary Table S1.

The 'GenTank' high-throughput thermocycler

A high-throughput water immersion-based thermocycler, dubbed 'GenTank', was constructed and employed to perform large numbers of PCR amplifications. A description of the system, including schematic diagrams, can be found in the Supplementary Data and online at http://genestamp.ibms.sinica.edu.tw/GenTank/index.htm. Briefly, the system consists of the following: three water tanks; a temperature-control module that sets the water temperatures of the three tanks to the appropriate levels for the various steps of PCR (i.e. denaturing, annealing and primer extension); and a robotic module that moves the PCR reaction vessels from one tank to the next. The system is capable of simultaneously performing thousands of PCR amplifications, or performing large-volume (milliliter-scale) PCR amplifications.

Capillary electrophoresis and data analysis

The generated PCR amplicons were diluted 1:20 in Hi-Di formamide solution (Applied Biosystems, Foster City, CA) containing a 1:100 dilution of GeneScan 2500 size standard (Applied Biosystems), and then separated using an Applied Biosystems 3100 genetic analyzer or 3730xl DNA analyzer. The capillaries were pre-run for 10 min before the samples were introduced by electrokinetic injection for 15 s at 15 kV. The electric field strength of the electrophoretic separations was set at 250 V/cm.

The resulting electropherograms were subjected to peak detection using a software program written in-house using Borland C++ builder 6.0 (Borland, Austin, TX). A detailed description of this software, along with the software itself, is available online at http://genestamp.ibms.sinica.edu.tw/AmpliconViewer/index.htm. The length of each PCR amplicon was determined by interpolation using curve-fit data for the electrophoresis size standard. First-order polynomial curve-fit data were used for amplicons <600 bp, while

second-order polynomial curve-fit data were used for larger amplicons (Supplementary Figure S4). The amount of each amplicon was semi-quantitatively determined based on the peak area.

For identification of transcript variants for a given gene, the amplicon lengths were compared with the lengths of the putative transcript variants archived in the ASTD (20) and Ensembl databases (21). Each variant in ASTD has a TRAN-prefixed transcript number, whereas those in the Ensembl database are designated by an ENST-prefixed transcript number. A given amplicon was either matched with the calculated length of an archived transcript variant, or it was regarded as representing a novel variant.

We then used a newly developed software program, which we called 'VariantAssembler', to identify long transcript variants that need to be resolved using multiple PCR primer pairs and amplifications. VariantAssembler was implemented in MATLAB (MathWorks, Natick, MA), and the program codes are available online at http://genestamp.ibms.sinica.edu.tw/VariantAssembler/index.htm. The program measures the similarities between the expression levels for the reference amplicons generated by the first PCR primer pair and the amplicons generated by additional PCR primer pairs, with the goal of identifying the variants expressed in the tested sample.

RESULTS

A high-throughput transcript variant analysis system

We constructed an integrated system for high-throughput analysis of transcript variants. The system is based on using PCR amplicon lengths to resolve putative transcript variants. Figure 1 depicts the procedural pipeline, which comprises six main components: (i) splice-site-flanking primer pair design; (ii) high-throughput oligonucleotide synthesis; (iii) high-throughput thermocycling for RT–PCR amplification; (iv) multi-channel capillary electrophoretic separation; (v) PCR amplicon analysis; and (vi) identification and quantification of transcript variants. The splice-site-flanking primer pairs were designed to flank the variable sequence regions of the putative transcript variants and generate PCR amplicons that each represented a distinct set of exon connectivity. Capillary electrophoresis was used to separate the amplicons, and the peak areas of the electropherogram were taken as representing the expression levels of the relevant amplicons. From there, the transcript variants were reconstructed based on the obtained information on exon connectivity and expression.

PCR primer pairs for transcript variant detection

A total of 16 715 genes and 93 441 putative transcript variants were downloaded from the ASTD database for analysis. Among the genes, 15 098 had multiple transcripts. To design splice-site-flanking primer pairs, each exon of the putative transcript variants was mapped to its genomic coordinates, and the common and discordant exons among the transcript variants of each gene were identified (Figure 2). To minimize the number of primer pairs needed for a given gene, we employed the optimal PCR primer flanking-region-selection scheme shown in Figure 2. As can be seen in the figure, a given transcript could be subject to various splicing events, such as exon insertion, the use of an ambiguous boundary, and others. The phrase 'ambiguous boundary' refers to either the start or end position of an exon when sequence information is inadequate to clearly identify one or both positions. Due to the amplicon length limit of capillary electrophoresis in achieving single-base separation, the primer pairs were designed to yield PCR amplicons shorter than 1 kbp. If the predicted size of the PCR amplicon was longer than 1 kbp, the PCR flanking region was split into smaller fragments that met this criterion (e.g. R2-1 and R2-2 in Figure 2).

The majority of the transcript variants obtained from ASTD were short EST fragments that contained only partial information on splicing events. Among the 93 441 transcript variants examined, 58% (54 221 of 93 441) were shorter than 800 bp, and 49 726 (53%) comprised fewer than five exons (Table 1). The primer design software successfully generated 22 307 PCR primer pairs for the 14 369 genes (1.48 primer pairs per gene on average). Among the 14 369 genes, 8680 required only one primer pair to flank all the variable splice sites, while 5689 required the use of two or more primer pairs.

The ASprimerDB web user interface

A web user interface (http://genestamp.sinica.edu.tw/ASprimer/index.htm) was established to provide detailed information on the splice-site-flanking primer pairs. Figure 3 shows a sample web page that provides information on the *VEGFB* gene. The query outputs include the Ensembl gene ID, the Entrez gene ID, the gene symbol and a brief description of the gene. The outputs also include web links to detailed information on each transcript, as found in the Ensembl and RefSeq databases. Two tables list information on the primers, including the ID of the primer pair (Primer pair ID), the forward/reverse primer sequences for each primer pair. The second table lists the predicted lengths of the PCR amplicons corresponding to the primer pair(s).

The transcript variants of *VEGFB* are displayed in the sample web page, along with the exon connectivity of each transcript variant and the size of each exon. Exons are shown as boxes and connected by solid lines. The transcript variants are color-coded to mark the different sources from which they were identified. The reference transcript is shown in cyan. The transcript variants obtained from the ASTD are shown in green and those transcripts other than the reference transcript from the Ensembl database are shown in blue. The transcript variants from the H-Invitational Database (H-InvDB) which provides curated annotation of human genes and transcripts are shown in red and have HIT-prefixed transcript number (22). The web page provides a view of the exon connectivity of each transcript variant, along with the PCR primer annealing sites. The number of exons and the starting and ending positions of each exon in the transcript variants of *VEFGB*, along with their sources, are presented in the table (bottom).

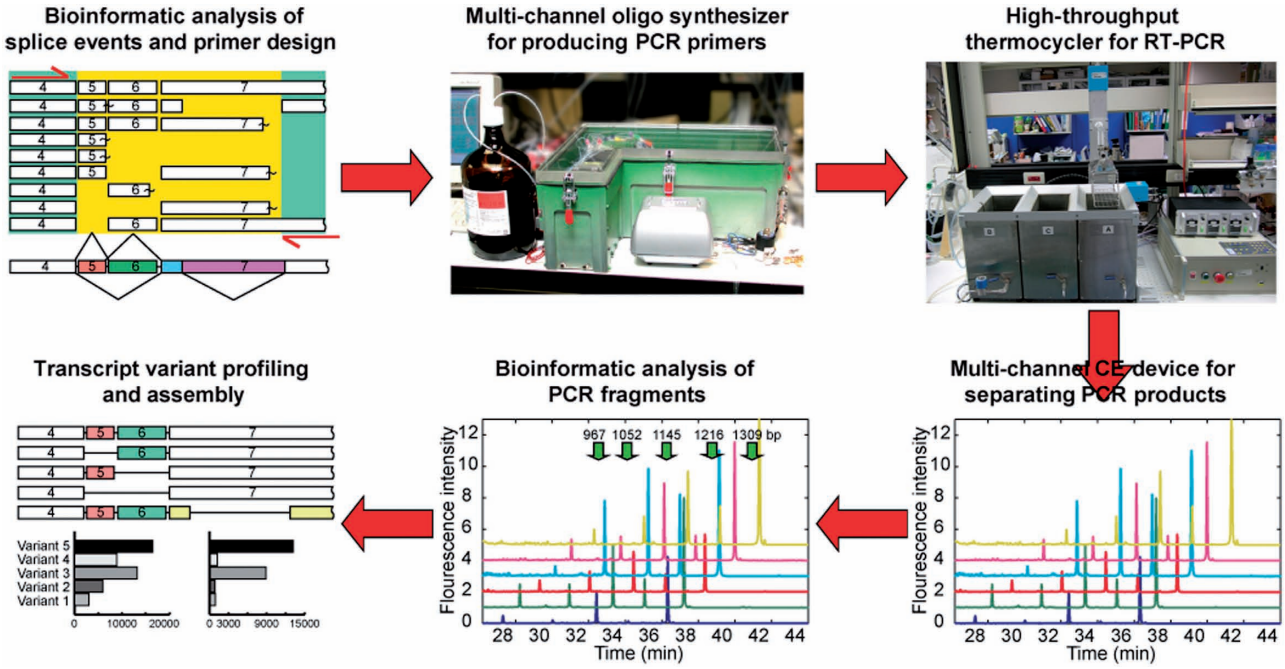

Figure 1. Flowchart of the high-throughput transcript variant profiling system. Putative splicing events are identified from databases and used to design splice-site-flanking primer sets. The primers are synthesized with a high-throughput oligonucleotide synthesizer and employed in a new high-throughput thermocycler, dubbed GenTank, to generate amplicons representing the transcript variants. The PCR amplicons are separated using a multi-channel capillary electrophoresis instrument, and the electropherograms are analyzed for semi-quantitative expression level measurement of each amplicon, and for assembly of transcript variants.

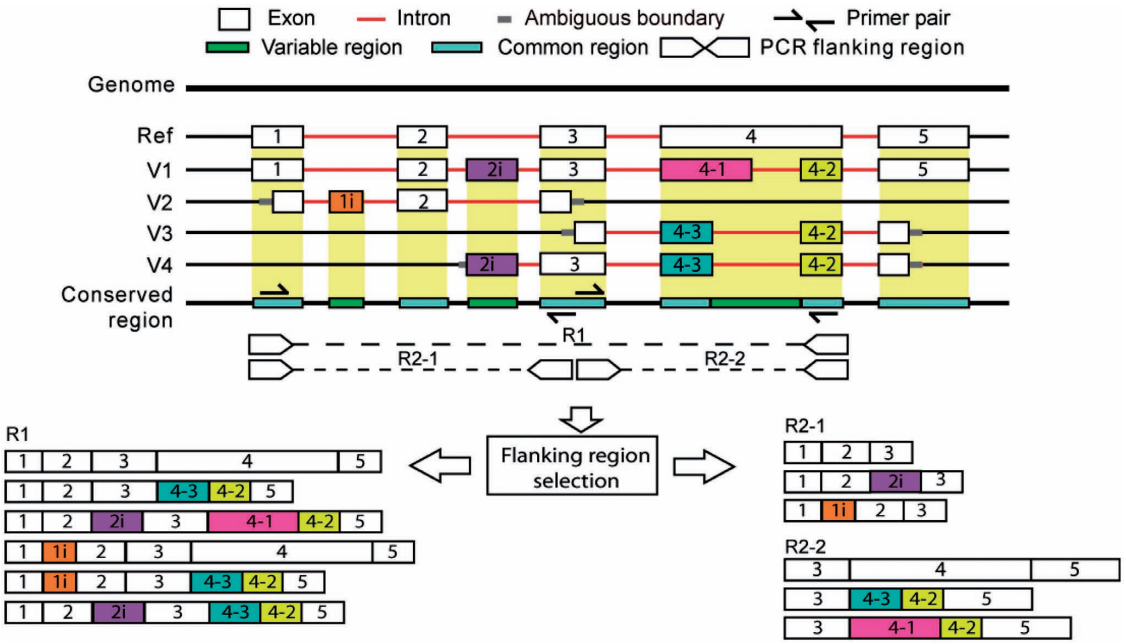

Figure 2. Schematic diagram illustrating the design of the splice-site-specific PCR primers. The reference transcript of a gene and its transcript variants documented in the two utilized databases are aligned with their genomic coordinates in order to define the common and variable exons present in the transcript variants. The canonical transcript of the Ensembl database is chosen as the reference transcript. The primers are located in consensus exon regions flanking the variable splice sites. The exon numbered with an 'i' suffix designates the insertion of an alternative exon. Numbered hyphenations indicate the presence of alternative splicing within the exons. For the sample transcripts shown in the figure, region R1 encompasses all the splicing events and can generate up to six possible transcript variants. If region R1 spans more than 1000 bp, then it is divided into two regions, R2-1 and R2-2, two primer pairs are used to flank the variable region.

Table 1. Summary of transcript variants and splice-site-flanking primers

Number of genes	16 715
Genes with multiple transcript variants	15 098
Genes with splice-site-flanking primers	14 369
Number of transcript variants	93 441
Variants shorter than 800 bp	54 211
Variants comprising less than five exons	49 726
Number of primer pairs	22 307
Number of genes requiring one splice-site-flanking primer pair	8680
Number of genes requiring two splice-site-flanking primer pairs	4075
Number of genes requiring three splice-site-flanking primer pairs	1189
Number of genes requiring more than three primer pairs	425
Average number of splice-site-flanking primer pairs per gene	1.48

We tested 278 genes with this system. The experimental results may be found in a Supplementary Data, in Excel format. We found that the results could be grouped into three categories. These categories are listed and described below, and selected examples are given to illustrate how the method resolves the exon connectivity of transcript variants.

Transcript variants resulting from one splicing event

The system identified numerous genes for which the expressed transcript variants arose from a single splicing event. Examples of this category include two genes known to be highly associated with cancer progression: *VEGFB* and *SPP1*. *VEGFB*, which was reported to play an important role in tumor angiogenesis, has three isoforms in human astrocytoma (23). *SPP1* (also known as *OPN*) is a ligand of the cell-surface receptors, integrin $\alpha_V\beta_3$ and CD44v3-6; it is up-regulated in breast cancer and contributes significantly to cancer cell invasion. Two transcript variants have been reported for *SPP1*, one with a deletion of exon 4 and the other with a deletion of exon 5 (24).

The ASTD database contained two transcript variants for *VEGFB*, resulting from an alternative 5′-donor splicing event located in exon 6. The Ensembl database contained only the transcript variant with a longer exon 6 (shown in Supplementary Data). The primer design software generated forward and reverse primers located in exons 1 and 7, respectively, to detect these splicing events. As show in Figure 4A, the electropherogram revealed that two peaks were obtained from the tested lung cancer specimen. Measurement of the elution times of the peaks allowed us to measure their sizes by curve fitting with respect to the size standards. Comparison with the sizes of the archived transcripts allowed us to determine that one variant corresponded to transcript ENST00000309422 in the Ensembl database, while the other corresponded to TRAN00000073236 (containing an alternative 5′ donor splicing event in exon 6) in ASTD.

The second example, *SPP1*, had five putative transcript variants listed in ASTD (see Supplementary Data); these putative variants arose from different types of splicing events, including exon insertion, exon deletion, an alternative 3′-acceptor splicing site and the use of intron retention. Some events had ambiguous boundaries because the transcript sequences were incomplete in ASTD. Four transcript variants were found in the Ensembl database; these arose from exon deletion and the use of intron retention. Only one transcript existed in both databases: ENST00000395080 in the Ensembl database corresponded to TRAN000000100841 in ASTD. For empirical validation of the putative transcript variants contained within the databases, our primer design software generated forward and reverse primer pairs in the consensus regions of exons 2 and 6, respectively. The experimental results revealed only two transcript variants from the tested lung cancer sample (Figure 4B); they corresponded to ENST00000395080 and ENST00000237623, and arose from exon 5 deletion splicing event.

Transcript variants resulting from multiple splicing events

The system was also capable of resolving transcript variants resulting from complex combinations of splicing events. Of the 278 genes that we tested for their transcript variants in lung adenocarcinoma tissue, eight yielded three or more transcript variants. One of these genes, *ABCC1* (also known as *MRP-1*), has been reported to have multiple transcript variants arising from complex combinations of splicing events in ovarian cancer, and some of these transcripts have been associated with multi-drug resistance in cancer (25). Figure 5A shows an electropherogram of the seven transcript variants our system identified for *ABCC1*, along with their peak areas (indicated beside each peak). Variants 1, 3 and 4 were found to be highly abundant, and the lowest versus highest expression levels differed by more than 30-fold.

The longest transcript variant, variant 1 (corresponding to ENST00000399410 in the Ensembl database), was used as the reference transcript for our splicing analysis. The results indicated that there were several possible splicing events among these variants, namely single or combinatorial deletions of exons 13, 16, 17 and 18. The transcript sizes seen in the electropherogram arose via deletions of exon 16, 17 and/or 18 (Figure 5A, yellow), with or without deletion of exon 13 (Figure 5A, blue). By matching the amplicon sizes with the predicted sizes of the different exon deletion combinations, we determined the exon connectivity of the transcript variants. For example, variants 6 and 7 both had deletions of exons 17 and 18 (Figure 5A, red arrows), but variant 7 had an additional exon 13 deletion. Variants 3 and 5 both had deletions of exon 17, but variant 5 also had a deletion of exon 13 (Figure 5A, orange arrows). The exon connectivity of variants 1, 2 and 4 are shown in the figure (Figure 5A, blue arrows for variants 1 and 2, turquoise arrow for variant 4). We did not observe any other exon deletion event in the *ABCC1* variants found in the tested sample.

To verify the accuracy of the exon connectivity computed by matching the sizes of the empirically determined amplicons with those of the putative transcript variants of *ABCC1*, we examined the expression levels of the transcript variants. First, the transcript variants were classified into two categories, namely with (Figure 5B, green column) or without (Figure 5B, blue column)

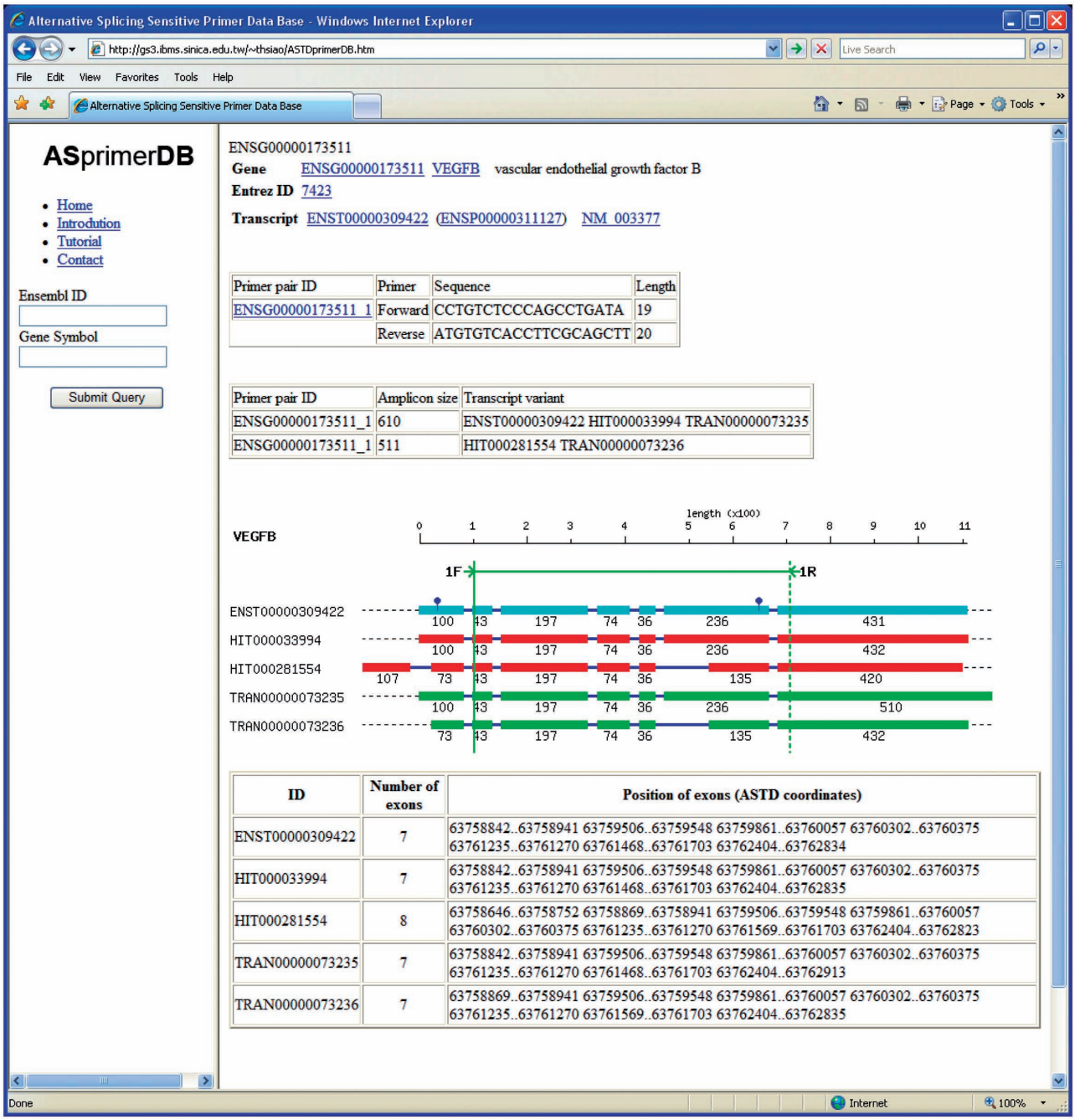

Figure 3. Browser view of the ASprimerDB database entry for the *VEGFB* gene. The web browser shows information on the splice-site-specific PCR primers, the exon connectivity of each transcript variant, the coordinates of the gene transcript, the lengths and locations of all exons, and the primer locations. The 3′-end positions of the forward and reverse primers are indicated by vertical green solid and dashed lines, respectively. The coordinates of each exon are listed at the bottom.

deletion of exon 13. Among the seven identified variants, the expression levels of the exon-13-deleted variants were consistently lower than those of variants that did not have a deletion of exon 13. Four splicing events with combinatorial deletions of exons 16, 17 and 18 (designated CSV1, CSV2, CSV3 and CSV4) were identified from the experimental data, and the transcript variants could be grouped based on these splicing events (Figure 5B). Analysis of the expression levels of the variants revealed that the transcript variants in which exon 13 was deleted could have the CSV1, CSV3, and CSV4 splicing events (Figure 5B, left). Variant 6 (V6) and variant 7 (V7) both had CSV1 splicing events that included deletion of exons 17 and 18. The CSV2 splicing event was only seen in transcript variants lacking exon 13. The only transcript variant to exhibit the CSV2 splicing event was transcript variant 4. The results from our expression analyses (Figure 5B) confirmed the exon connectivity determined by our system

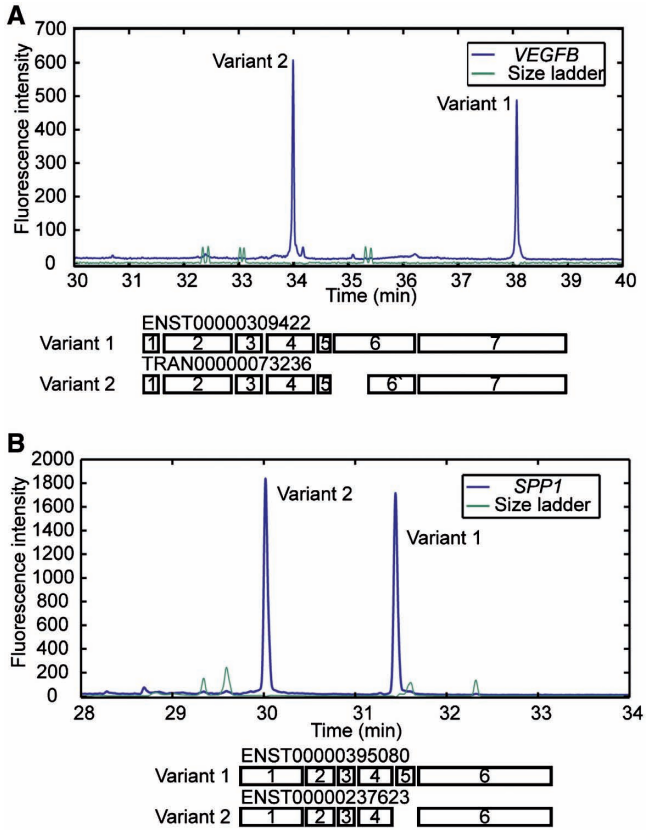

Figure 4. Transcript variant analysis of *VEGFB* and *SPP1*. Every peak in the electropherogram reflects a PCR amplicon comprising a stretch of consecutive exons. The size of the amplicon is calculated with respect to the curve-fitted data of the size standard. (**A**) Transcript variant analysis of the *VEGFB* gene. Two transcript variants were detected and found to differ by an alternative 5′ donor event; they were identified as corresponding to ENST00000309422 and TRAN00000073236 in the Ensembl and ASTD databases, respectively. (**B**) Transcript variant analysis of the *SPP1* gene. Two variants that differ by an exon 5 deletion event were detected.

(Figure 5A). Among the seven transcript variants detected for *ABCC1*, variant 4 (with deletions of exons 16 and 17) and variant 5 (with deletions of exons 13 and 17) were reported in the literature but not documented in both the Ensembl and ASTD databases. Variant 7 (with deletions of exons 13, 17 and 18) is a novel variant that had not previously been reported in either the literature or the Ensembl and ASTD databases.

Collectively, these results show that our system can successfully detect and resolve the exon connectivities of transcript variants based on size matching and expression level verification.

Transcript variants that must be obtained using multiple PCR primer pairs

Capillary gel electrophoresis based DNA sequencing is the currently available technology for resolving single base difference in DNA fragments as large as 1 kb. To distinguish transcript variants which have predicted amplicon size larger than 1 kbp, it is necessary to use multiple primer pairs to amplify the transcript variants for analysis. The exon connectivity results are then concatenated to assemble the full-length transcript.

ITGB4 is a long gene whose transcripts are longer than 5.8 kbp. We found two transcript variants, ENST00000200181 and ENST00000339591, in the Ensembl database; and three transcript variants, TRAN 00000077450, TRAN00000077458 and TRAN 00000077454, in ASTD (Figure 6A). ENST00000200181 had an exon 33 deletion, while ENST00000339591 had an exon 35 deletion; both have been described in the literature (6). TRAN00000077454 and TRAN00000077458 contained ambiguous exon boundaries and lacked complete sequence information in ASTD. The gene has a relatively long variable region, meaning that the use of a single PCR primer pair would generate long amplicons that would be difficult to separate by high-resolution capillary electrophoresis. Thus, we designed two primer pairs: primer pair 1 was located in exons 17 and 27 while primer pair 2 was located in exons 28 and 37 to flank the variable regions (Figure 6A).

We then used our system to elucidate the variants of this gene in the tested lung cancer sample. As shown in Figure 6B, primer pair 1 yielded amplicons 1 and 2, the latter of which was shorter and encompassed the exon 18 deletion event documented in the TRAN00000077458 transcript variant (ASTD database). Primer pair 2 yielded three PCR amplicons reflecting deletions of exons 33, 35 and 36. Based on analysis of the amplicon lengths (Figure 6B) and exon connectivity, ENST00000339591 was not detected in the tested sample, but amplicon 3 (with deletion of exon 33) was consistent with the computed length of ENST 00000200181, amplicon 4 corresponded to TRAN 00000077450, and amplicon 5 most likely corresponded to TRAN00000077454. Although the sequence information from ASTD was incomplete, primer pairs 1 and 2 appeared to amplify TRAN00000077458 and TRAN 00000077454, respectively. Regarding the expression levels of the various *ITGB4* transcript variants in the lung cancer tissue specimen, primer pair 1 generated 94.3 and 5.7% of amplicons 1 and 2, respectively, while the proportion of amplicons 3, 4 and 5 generated by primer pair 2 was 2.9, 90.6 and 6.5%, respectively (Figure 6C).

To decipher the exon connectivity, we permuted combinations of consecutive exons based on the possible splicing events embodied by amplicons obtained from the two PCR amplifications (Figure 6D). Variants 1, 2 and 3 could be generated by connecting amplicon 1 with amplicons 3, 4 and 5, respectively, whereas variants 4, 5 and 6 were could be generated by connecting amplicon 2 with amplicons 3, 4 and 5, respectively. However, although six putative transcript variants could be constructed, not all of these variants were expressed in the tested sample. Based on the idea that the two amplicons representing a single transcript variant should have similar proportions in the PCR products (following normalization across the different RT-PCR amplifications), we used similarities in expression level to pair the amplicons into potential sets of expressed variants.

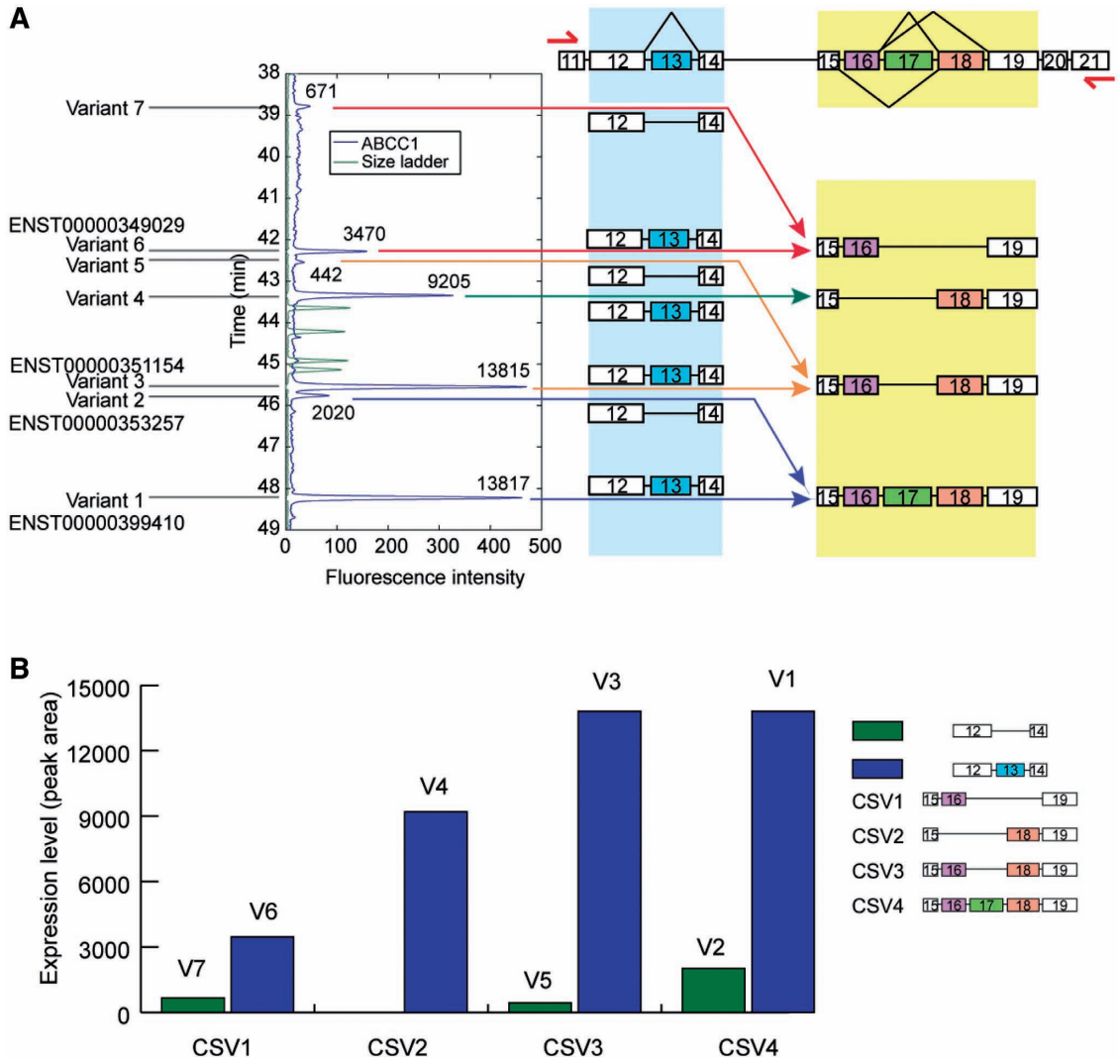

Figure 5. Transcript variant analysis of the *ABCC1* gene, which has multiple splicing events. (**A**) Seven transcript variants of *ABCC1* were detected. The transcript variants documented in the Ensembl database are marked with the Ensembl accession number. The detected transcript variants are shown by the blue electropherogram and the size ladders are shown by the green electropherogram. The colored arrows indicate the exon connectivity of each transcript variant detected in the electropherogram. Based on informatic analysis, we determined that the *ABCC1* variants arose from an alternative exon 13 deletion (blue) and combinations of splicing events (CSVs; here, deletions) among exons 16, 17 and 18 (yellow). The under-peak areas are shown beside each peak. (**B**) Analysis of the transcript variants with alternative exon 13 deletion and the CSVs. Based on the alternative exon 13 deletion, the 7 variants can be grouped into 4 CSVs. The exon 13-deleted variants were all expressed at low levels.

A mathematical formula was used to measure expressional similarities between the empirical electropherogram data and putative transcript variants composed of different amplicons, in order to identify the most probable putative transcript variants that were expressed in the test sample, as follows:

$$S(x_1,x_2) = \begin{cases} 1 - 10 \times |x_1 - x_2| , & \text{if} \quad |x_1 - x_2| < 0.1, \\ 0 , & \text{if} \quad |x_1 - x_2| \geq 0.1, \end{cases}$$

where x_1 and x_2 are the product percentages represented by the amplicons arising from primer pairs 1 and 2, respectively. More specifically, x_1 is the percentage of the first amplicon in the first PCR reaction, and x_2 is the summed percentages of the amplicons in the second PCR reaction that are connected to the amplicon represented by x_1. Using Figure 6D as an example, we see that for transcript set S1, x_1 is 94.3% (which is the percentage of amplicon 1 in the results from the first PCR) and x_2 is 93.5% (which is the summed percentages of amplicons 3 and 4, which are connected with amplicon 1 to constitute transcript variants *V*1 and *V*2, respectively). The measurement is based on the principle that the similarity score, $S(x_1, x_2)$, is close to 1 when x_1 and x_2 are 'similar', i.e. when the amplicons constituting the transcript variants have similar proportions in the products generated by the two PCR amplifications. The transcript variants contained within the transcript set having the highest similarity score are considered to be expressed.

This calculation assumes that the number of expressed transcripts does not exceed the maximum number of amplicons detected in the PCR amplifications. In the

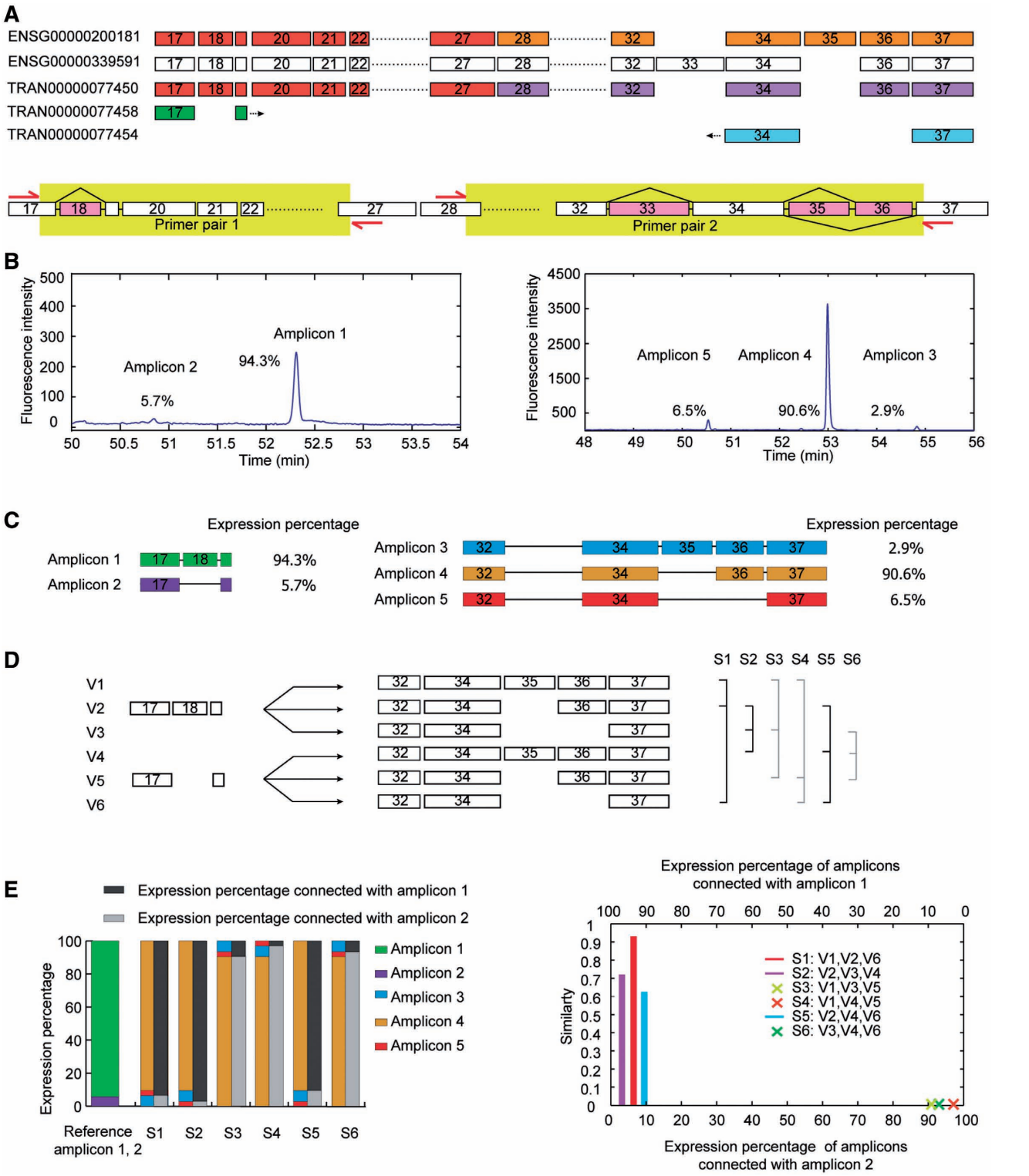

Figure 6. Transcript variant analysis of *ITGB4*, which was amplified by two primer pairs. (**A**) The *ITGB4* gene structure and alignment of the transcript variants documented in the databases. Two exon stretches (yellow) were amplified to reveal the complete transcript variants of this long (5.8 kbp) gene. (**B**) Electropherograms of the amplicons derived from these two PCR amplifications using DNA from a lung cancer specimen. The first amplification generated amplicons 1 and 2, while the second generated amplicons 3, 4 and 5. The expression percentages for each of these PCR amplicons are denoted in the figures. (**C**) The exon stretches encompassed by each amplicon, and their expression percentages in the products of the two PCR amplifications. (**D**) Six putative variants were assembled by permuting the exon stretches generated by the two PCR amplification reactions. Six transcript sets (S1–S6), each consisting of three putative transcript variants, were composed and used to identify which transcript variants could account for the peaks observed in the electropherograms. (**E**) Identification of the transcript variants expressed in the test sample. The left panel shows an expression pattern comparison between reference amplicons 1 and 2 and the expression patterns of the transcript sets. The colored bars indicate the expression percentages of each amplicon. The dark and light gray bars indicate the expression percentages of the amplicons from the second PCR in association with those of amplicons 1 and 2, respectively. An expression pattern similarity score was calculated for each transcript set; the results are shown on the right. The upper abscissa indicates the expression percentage of the amplicons from the second PCR connected with amplicon 1, while the lower abscissa indicates the expression percentage connected with amplicon 2. The ordinate indicates the similarity score.

case of *ITGB4*, a maximum of three amplicons were generated (in this case, by primer pair 2). Six possible expressed transcript sets consisting of three transcript variants each (labeled S1–S6) were generated from the putative transcript variants shown in Figure 6D. Taking transcript set S1 (consisting of V1, V2 and V6) as an example, we see that in the first PCR reaction both V1 and V2 would yield amplicon 1, which comprised 94.3% of the empirical yield, whereas V6 would yield amplicon 2, which comprised 5.7% of the PCR product generated by primer pair 1. In the second PCR amplification, V1 and the V2 would yield amplicons 3 and 4, respectively, which together contributed 93.5% of the yield. V6 was calculated separately from the above two variants because it was connected with amplicon 2; in this case, the second PCR reaction would yield amplicon 5, which comprised 6.5% of the generated PCR product.

We then performed this calculation for all six possible transcript sets. As shown in Figure 6E (left panel), transcript sets S1, S2 and S5 have summed expression percentages that are fairly similar to the empirical results for amplicon 1 (dark gray) and amplicon 2 (light gray). Computation of similarity scores for the expression patterns manifested by six transcript sets using the above formula yielded similarity scores for S1, S2 and S5 of 0.92, 0.72 and 0.63, respectively. Based on this score, we concluded that transcript variants V1, V2 and V6 (corresponding to set S1) of *ITGB4* were expressed in the tested tumor specimen.

DISCUSSION

This report describes a novel integrated system for empirically detecting and verifying putative transcript variants. PCR primers that flank the variable regions of transcript variants (and thus encompass the possible splicing events) are used to amplify the variants, and a high-resolution capillary-based DNA sequencer is used to discriminate among transcript variants whose sequences may differ by as little as a single base. The system was able to detect two expressed transcript variants for each of two exemplar genes (*VEGFB* and *SPP1*), and correctly identify these variants from the ASTD and Ensembl databases. In addition, the system could detect two different types of alternative splicing event within these two genes (e.g. an alternative 5′ donor site and exon deletion) based solely on the amplicon length. Thus, the system allows simultaneous detection of various splicing patterns among expressed transcript variants.

Although the primer design described herein was based on documented transcript variants, this system is not limited to known variants. On the contrary, the system can detect novel transcript variants, as demonstrated by the identification of a novel transcript variant of *ABCC1*. The transcript variants of *ABCC1* were covered using a single primer pair encompassing combinatorial splicing events (i.e. various deletions of exons 13, 16, 17 and 18). By considering both the sizes and quantities of PCR product represented by the various amplicons, the system effectively resolved the complex exon connectivity represented by the transcripts of *ABCC1* and determined that seven transcript variants were expressed in the tested sample.

Rather than detecting the expressional variations among individual exons, as is done in exon microarray experiments, our system measures the expressional variations among consecutive stretches of exons. We found that the transcript variants of ~60% (8680/14 369) of the examined genes could be resolved using a single PCR primer pair, and 89% (12 755/14 369) genes could be resolved using one or two primer pairs. On average, only 1.48 primers were required per gene. In cases for which a single primer pair was sufficient, each amplicon peak represented a single transcript variant, and the peak area could be quantitatively estimated as the expression level. For more complex cases, such as that of *ITGB4*, the expression patterns of consecutive exon stretches were aligned to assemble the full array of transcript variants, and expression pattern similarity scores were computed for each transcript set and used to identify the expressed variants.

The exon connectivity of each gene transcript determines whether the encoded protein contains a frame shift or premature stop codon, whether it could be vulnerable to nonsense-mediated decay, and/or if it contains changes that affect the functional domains of the protein. Bioinformatic analysis of EST libraries can provide an initial blueprint of splicing events, but it is difficult to resolve the exon connectivity of a full transcript from those based on EST sequences, which are typically short. Some studies have listed all of the possible transcript variants that may be generated from EST alignment (26), but consideration of all possible combinations of splice events generates a huge number of putative transcript variants. Several other studies have applied constraints, for example by setting a threshold to limit the number of the aligned EST or mRNA sequences (5,27). In addition, probability models have been employed to calculate maximum likelihoods from the aligned EST sets, in order to identify transcript variants that have higher possibilities of being expressed (12).

Our method improves on these previous efforts by integrating high-throughput *in silico* and empirical approaches. The empirical verification minimizes the number of putative transcript variants and identifies those that are truly expressed in a given sample. It does so by considering both the connectivity of the amplicons and their expression patterns in the tested sample. Expression pattern similarity based on the abundance of variant transcripts is important in determining which expression pattern (transcript set) is most likely to be real. However, the abundance of variants is not the only relevant parameter. Our method assumes a maximum number of transcript variants based on a count of amplicon peaks in the electropherogram. A simulation is performed to assess the accuracy of this assumption, using Bayesian probability theory. The transcript variants forming the expression pattern with the highest similarity score are identified as expressed (see the Supplementary Data for more details). In the case of *ITGB4*, for example, the computation assumed that there was a maximum of

three transcript variants in an expression pattern, and the simulation showed that the accuracy of this prediction was 97.6%. Although it is formally possible that *ITGB4* had more than three transcript variants in the tested sample, the empirical electropherogram data show that the expression of any additional transcript variant was insignificant. To verify this contention, we determined the detection limit of the system. In the case of *ITGB4* of Figure 6B, the signal-to-noise (S/N) ratio of amplicon 4 was 931. By setting the detection limit at an S/N ratio of 3, the limit was thus 1/310 (0.3%) for this most abundant amplicon of *ITGB4*.

Figure 6 demonstrates how the VariantAssembler program assembles and identifies the transcript variants expressed in a test sample. Long transcript variants requiring three or more primer pairs are assembled through serial concatenation of the variant assembly results. As shown in Supplementary Figure S6, our method showed that V1, V2 and V6 of *ITGB4* were expressed in the test sample, and further indicated the expression level of each transcript variant. VariantAssembler uses the transcript variants and their expression levels identified from the first and second PCRs (e.g. the V1, V2 and V6 in Supplementary Figure S6) as the x_1 inputs. In cases where a third primer pair is used, VariantAssembler uses the amplicons generated by the third PCR, along with their quantities (taken from the electropherogram), as inputs for x_2. By serially inputting the information from the various PCR amplifications, VariantAssembler can resolve the full-length transcript variants expressed in a given test sample.

It should be noted that the method does not determine the exact transcript variants present in a test sample under an extreme condition, which is all amplicons have the same value of expression percentage and the number of amplicons generated by every primer pair in a multi-primer pair amplification is the same. For examples, 50/50 for two amplicons generated by every primer pair or 33/33/33 for three amplicons generated by every primer pair in the amplification set (see the Supplementary Data for more details). Long-range PCR amplification and sequencing would be the way to resolve the above situation. However, with a detection limit of 0.3% expression percentage as determined in Figure 6, the situation of having amplicons with indistinguishably same expression percentage is rare.

Two splicing event types are not included in this report: the use of alternative transcript start sites, and alternative polyadenylation. These two types of alternative splicing events generate transcript variants that lack the regions required to design 5′- or 3′-end primers to flank the variable region for amplification. These two types of splicing events are solved by designing one primer in the common region and a specific primer for each putative transcript variant.

The empirical exon microarray experiments have been widely used to measure the expression of exons and detect splicing events. However, it has proven difficult to verify these transcript-specific expression results. For example, qPCR is not adequate to resolve the amplicons, and RT–PCR followed by agarose gel electrophoresis is time consuming and labor intensive. On the other hand, high throughput sequencing using NGS methods generates millions of short reads with lengths of a few hundred bases at best. As with an exon microarray, the NGS methods detect splicing events and identify putative transcript variants that need to be verified for correct assembly.

To fulfil verification goals, the system has to be able to detect transcript variants that are incompletely annotated in the databases. A simulation was performed to verify the robustness of the method when dealing with incomplete annotation. Exons were randomly removed from the annotation. The results are shown in a supplemental data file (Supplementary Figure S8). In brief, by using the *SPP1, VEGFB, ABCC1* and *ITGB4* genes discussed in the present report as examples, the results show that our method can identify novel splicing events or exons by examining length discrepancies between empirical and predicted PCR amplicons.

The system is also able to detect novel splicing events other than exon-skipping or deletion because such events are revealed by length discrepancies between empirically verified transcript variants and variants annotated in the databases. However, to confirm or assemble putative variant transcripts identified by methods other than the NGS approach (e.g. using RT–PCR), additional sequence information may be needed to pinpoint novel splicing events. The enhanced sequence information provided by the NGS method aids in the resolution of transcript variants resulting from mutually exclusive exon splicing events within a gene, which might yield amplicons of the same length. To estimate the probability of encountering such a situation, we analyzed putative splicing events in the ASTD database that would generate amplicons of the same length for any particular gene, and found that this scenario is rare. Of 93 441 putative transcript variants derived from 16 715 genes using 78 165 putative splicing events in the ASTD database, 49 genes would yield two transcript variants embodied by amplicons of the same length because of the presence of 56 mutually exclusive exon splicing events (0.07%), and 22 genes (0.13%) would yield two transcripts embodied by amplicons of the same length because two independent exon-skipping events would be expected. These genes are tabulated in a Supplemental Data, in Excel format. Furthermore, as it is known that different tissues may have different expressed transcript variants, the probabilities of encountering the scenarios outlined above are much less than the calculated percentages when transcript variants of a gene are studied in any particular type of tissue. Therefore, our method is applicable to more than 99% of genes in the ASTD database, and accurately verifies putative transcript variants suggested by methods that lack sequence information on splicing events.

Our new method provides a new means to rapidly identify and/or validate the differentially expressed transcript variants revealed by exon microarray and NGS data. The PCR primer pairs that may be used for the genes currently known to have transcript variants are archived in the database developed for this study, and

capillary electrophoresis-based sequencers are widely available. Although the GenTank high-throughput thermocycler is not commercially available, its working principle is straightforward and information on constructing the instrument is provided in the supplementary file.

This report not only describes a pipeline of experimental procedures and instruments, it also provides the utilized software. Users can visualize the primer locations, transcript intron–exon structures, exon connectivity, exon lengths and other information using the ASprimerDB database and its web interface. The peak-detection program, AmpliconViewer, may be used to analyze the peak location and peak area in the electropherogram. The VariantAssembler program may be used to calculate expression pattern similarity scores and identify the expressed versus putative transcript variants for genes with long transcript variants or complex splicing events. These tools will greatly facilitate the work of researchers who are interested in studying transcript variants. Our system is capable of performing more than 2000 reactions in a day. Using integrated technologies that combine primer design, high-throughput oligonucleotide synthesis, high-throughput GenTank thermocycling, capillary electrophoresis-based DNA sequencing and data-processing software, our novel system can be used to comprehensively analyze the alternative splicing patterns of thousands of genes in a short time.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

The Academia Sinica Thematic Project (AS-94-TP-B02); NRPGM grants from the Department of Health (DOH97-TD-G-111-025, DOH98-TD-G-111-014); the National Science Council of Taiwan, ROC (NSC-93-3112-B-001-013-Y). Funding for open access charge: Academia Sinica, Taiwan, ROC.

Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. Shkreta,L., Froehlich,U., Paquet,E.R., Toutant,J., Elela,S.A. and Chabot,B. (2008) Anticancer drugs affect the alternative splicing of Bcl-x and other human apoptotic genes. *Mol. Cancer Ther.*, **7**, 1398–1409.
3. Lee,C., Atanelov,L., Modrek,B. and Xing,Y. (2003) ASAP: the alternative splicing annotation project. *Nucleic Acids Res.*, **31**, 101–105.
4. Thierry-Mieg,D. and Thierry-Mieg,J. (2006) AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol.*, **7 (Suppl. 1)**, S12 11–14.
5. Kim,N., Shin,S. and Lee,S. (2005) ECgene: genome-based EST clustering and gene modeling for alternative splicing. *Genome Res.*, **15**, 566–576.
6. Gardina,P.J., Clark,T.A., Shimada,B., Staples,M.K., Yang,Q., Veitch,J., Schweitzer,A., Awad,T., Sugnet,C., Dee,S. *et al.* (2006) Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC Genomics*, **7**, 325.
7. Xi,L., Feber,A., Gupta,V., Wu,M., Bergemann,A.D., Landreneau,R.J., Litle,V.R., Pennathur,A., Luketich,J.D. and Godfrey,T.E. (2008) Whole genome exon arrays identify differential expression of alternatively spliced, cancer-related genes in lung cancer. *Nucleic Acids Res*, **36**, 6535–47.
8. 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.
9. Wilhelm,B.T., Marguerat,S., Watt,S., Schubert,F., Wood,V., Goodhead,I., Penkett,C.J., Rogers,J. and Bahler,J. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, **453**, 1239–1243.
10. Sultan,M., Schulz,M.H., Richard,H., Magen,A., Klingenhoff,A., Scherf,M., Seifert,M., Borodina,T., Soldatov,A., Parkhomchuk,D. *et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**, 956–960.
11. Xing,Y. and Lee,C. (2008) Reconstruction of full-length isoforms from splice graphs. *Methods Mol. Biol.*, **452**, 199–205.
12. Xing,Y., Yu,T., Wu,Y.N., Roy,M., Kim,J. and Lee,C. (2006) An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. *Nucleic Acids Res.*, **34**, 3150–3160.
13. Anton,M.A., Gorostiaga,D., Guruceaga,E., Segura,V., Carmona-Saez,P., Pascual-Montano,A., Pio,R., Montuenga,L.M. and Rubio,A. (2008) SPACE: an algorithm to predict and quantify alternatively spliced isoforms using microarrays. *Genome Biol.*, **9**, R46.
14. Wang,H., Hubbell,E., Hu,J.S., Mei,G., Cline,M., Lu,G., Clark,T., Siani-Rose,M.A., Ares,M., Kulp,D.C. *et al.* (2003) Gene structure-based splice variant deconvolution using a microarray platform. *Bioinformatics*, **19 (Suppl. 1)**, i315–i322.
15. Shai,O., Morris,Q.D., Blencowe,B.J. and Frey,B.J. (2006) Inferring global levels of alternative splicing isoforms using a generative model of microarray data. *Bioinformatics*, **22**, 606–613.
16. Fehlbaum,P., Guihal,C., Bracco,L. and Cochet,O. (2005) A microarray configuration to quantify expression levels and relative abundance of splice variants. *Nucleic Acids Res.*, **33**, e47.
17. Klinck,R., Bramard,A., Inkel,L., Dufresne-Martin,G., Gervais-Bird,J., Madden,R., Paquet,E.R., Koh,C., Venables,J.P., Prinos,P. *et al.* (2008) Multiple alternative splicing markers for ovarian cancer. *Cancer Res.*, **68**, 657–663.
18. Schindler,S., Heiner,M., Platzer,M. and Szafranski,K. (2009) Comparison of methods for quantification of subtle splice variants. *Electrophoresis*, **30**, 3674–3681.
19. Cheng,J.Y., Chen,H.H., Kao,Y.S., Kao,W.C. and Peck,K. (2002) High throughput parallel synthesis of oligonucleotides with 1536 channel synthesizer. *Nucleic Acids Res.*, **30**, e93.
20. Koscielny,G., Le Texier,V., Gopalakrishnan,C., Kumanduri,V., Riethoven,J.J., Nardone,F., Stanley,E., Fallsehr,C., Hofmann,O., Kull,M. *et al.* (2009) ASTD: the alternative splicing and transcript diversity database. *Genomics*, **93**, 213–220.
21. Flicek,P., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F., Cutts,T. *et al.* (2008) Ensembl 2008. *Nucleic Acids Res.*, **36**, D707–D714.
22. 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.
23. Gollmer,J.C., Ladoux,A., Gioanni,J., Paquis,P., Dubreuil,A., Chatel,M. and Frelin,C. (2000) Expression of vascular endothelial growth factor-b in human astrocytoma. *Neuro Oncol.*, **2**, 80–86.

24. He,B., Mirza,M. and Weber,G.F. (2006) An osteopontin splice variant induces anchorage independence in human breast cancer cells. *Oncogene*, **25**, 2192–2202.
25. He,X., Ee,P.L., Coon,J.S. and Beck,W.T. (2004) Alternative splicing of the multidrug resistance protein 1/ATP binding cassette transporter subfamily gene in ovarian cancer creates functional splice variants and is associated with increased expression of the splicing factors PTB and SRp20. *Clin. Cancer Res.*, **10**, 4652–4660.
26. Leipzig,J., Pevzner,P. and Heber,S. (2004) The alternative splicing gallery (ASG): bridging the gap between genome and transcriptome. *Nucleic Acids Res.*, **32**, 3977–3983.
27. Xing,Y., Resch,A. and Lee,C. (2004) The multiassembly problem: reconstructing multiple transcript isoforms from EST fragment mixtures. *Genome Res.*, **14**, 426–441.