Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Mining ORESTES no-match database: can we still contribute to cancer transcriptome?

Rogério da Silva Fonseca[1], Dirce Maria Carraro[2] and Helena Brentani[1]

[1]Laboratório de Bioinformática, Hospital do Câncer, São Paulo, SP, Brasil
[2]Laboratório de Análise de Expressão Gênica,
Instituto Ludwig para Pesquisa sobre o Câncer, São Paulo, SP, Brasil
Corresponding author: H. Brentani
Email: helena@lbc.ludwig.org.br

Genet. Mol. Res. 5 (1): 24-32 (2006)
Received September 22, 2005
Accepted January 30, 2006
Published February 24, 2006

ABSTRACT. The Human Cancer Genome Project generated about 1 million expressed sequence tags by the ORESTES method, principally with the aim of obtaining data from cancer. Of this total, 341,680 showed no similarity with sequences in the public transcript databases, referred to as "no-match". Some of them represent low abundance or difficult to detect human transcripts, but part of these sequences represent genomic contamination or immature mRNA. We performed a bioinformatics pipeline to determine the novelty of ORESTES "no-match" datasets from prostate or breast tissues. We started with 14,908 clusters mapped on the human genome. A total of 2226 clusters originating from more than two libraries or singletons with gaps upon genome alignment were selected. Ninety-four clusters with canonical splice sites representing the most stringent criteria to be considered a gene were subjected to manual inspection regarding genomic hits. Of the manually inspected clusters, 49.6% contained new sequences where 42.2% were probable low-expression alternative forms of the characterized genes and 7.4% unpredicted genes. RT-PCR followed by sequencing was performed to validate the largest spliced sequence from 8 clusters, resulting in the confir-

 ©FUNPEC-RP www.funpecrp.com.br

mation of five sequences as true human transcript fragments. Some of them were differentially expressed between tumor and normal tissue by an *in silico* analysis. We can conclude that after clean up of the no-match dataset, we still have about 939 new exons and 165 unpredicted genes that could complete the prostate or breast transcriptome.

Key words: Gene discovery, ORESTES, Expressed sequence tags, Cancer, Bioinformatics

INTRODUCTION

To understand the genetic basis of an organism, one of the main steps is the determination of all genes in its genome. To do so, two different approaches have been successfully applied: gene prediction by software (Claverie, 1997) and generation of short cDNA sequences, expressed sequence tags (ESTs) (Adams et al., 1991). The accuracy of prediction programs for the human genome is about 70%, but they generally found to be inefficient in the identification of small exons and also in finding genes that escape known patterns (Lander et al., 2001). ESTs, the second approach, relies on direct experimental data, but is subject to artifacts arising from immature mRNA and sequence contamination by genomic DNA, nucleic acids from other organisms (e.g., bacteria) and chimeric DNAs (Sorek and Safer, 2003). Nevertheless, ESTs represent transcript diversity in terms of the expressed genes, and their transcript forms (Wolfsberg and Landsman, 1997; Modrek et al., 2001; Wang et al., 2003) have been helpful in gene discovery and characterization (Adams et al., 1992; Liang et al., 2000; Ferreira et al., 2004), in the study of expression patterns (Vasmatzis et al., 1998; Bortoluzzi et al., 2000; Yu et al., 2001), and in gene physical mapping (Hudson et al., 1995).

The aim of the Human Cancer Genome Project (LICR/FAPESP-HCGP) was to contribute to the annotation of genes in the human genome (de Souza et al., 2000) and to obtain data from cancer, principally to identify cancer-related genes (Brentani et al., 2003). This project yielded 1,190,044 ESTs using the ORESTES (open-reading frame ESTs) method (Camargo et al., 2001), which is highly sensitive, capable of detecting low-abundant transcripts. ORESTES has a representation bias towards the central region of the transcript (Dias Neto et al., 2000), distinguishing it from most EST-generating methods, which favor instead the 5' or 3' regions of the cDNA, usually associated with non-coding sequences (Okubo et al., 1992). Otherwise this method uses no specific poly-T primers for cDNA syntheses favoring genomic amplification. From all ORESTES generated in this project, 341,680 showed no similarity with other sequences in the public transcript databases, as evaluated in 2000 (Camargo et al., 2001). These sequences were called "no-match" and were ESTs representing low abundance or difficult to detect transcripts; therefore, have not been identified to date by other experimental approaches. They can also represent genomic contamination or DNA or RNA from contaminating microorganisms. To determine the unexplored potential of ORESTES no-matches, we reevaluated a subset of these no-match sequences derived from two tissues with high cancer incidence in Brazil, prostate and breast (www.inca.gov.br).

MATERIAL AND METHODS

Computational methods

HCGP databank contains all the ORESTES generated in the Human Cancer Genome Project, annotated according to hits to public transcript databanks. The sequences that were annotated as no-match in HCGP were searched in map4 databank. The latter (July, 2002) contains all cDNA available in dbEST (Boguski et al., 1993) and mRNA sequences from known human genes from UniGene release 153 (Schuler et al., 1996) in clusters based on genomic alignment (masked human genome release 29, from NCBI). The programs and criteria for clustering are described elsewhere (Sakabe et al., 2003 and Galante et al., 2004).

Some sequences derived from libraries containing genomic DNA or immature mRNA contamination were retired (Sorek and Safer, 2003). Both databases, HCGP and map4, are kept in the Computational Biology Laboratory of the Ludwig Institute, São Paulo Branch, and the computational procedures used to mine them were based on PERL scripts.

The only clusters selected were those that contained at least one no-match sequence derived from prostate or breast tissues and that were formed by ESTs originating from at least two distinct libraries. The rationale is that the larger the numbers of library present in a cluster, the lower the probability of the cluster comprising contaminating sequences. Nevertheless, singletons that show gaps upon genomic alignment were not removed. The removal of clusters that align with full-length transcripts or ESTs (UniGene build 160) of other projects in the map4 databank was carried out to update the preceding analysis (Camargo et al., 2001). The selection of clusters with sequences that show gaps upon their alignment with the genome, which suggest splicing events characteristic of eukaryotic transcripts, and clusters with canonical splice sites (GT…AG), was manually evaluated. This splice site search was performed considering an extra 5-bp upstream and 5-bp downstream of the internal extremities of the aligned segments. This safety margin was adopted to avoid false-negative results due to sequence displacement, which may occur by the use of the heuristic BLAST alignment tool (Burset et al., 2000). Manual evaluation was performed in NCBI Map Viewer and the UCSC genome browser (Kent et al., 2002). Some sequences with confirmed gapped hits in regions that contained *ab initio* predicted genes in MapViewer proceeded to experimental validation.

Experimental validation

For confirmation that the selected ORESTES derived from human transcripts, experimental validation with cDNA of the same type of tissue from their original dataset was performed. Total RNA was extracted with Trizol (Invitrogen) from excised and micro-dissected human breast tumor, and normal and tumor human prostate samples, following the manufacturer's instructions. The RNA quality was evaluated by electrophoresis on 1% agarose/TBE gels. Samples with intact RNA were treated with DNAseI (Invitrogen) (1 U/2.5 µg total RNA), according to the manufacturer's instructions, and the absence of genomic DNA was confirmed by PCR using primers for introns 12 and 13 of the human mut-L homologue 1 gene (hMLH1) - (forward - 5' TGGTGTCTCTAGTTCTGG 3' and reverse - 5' CATTGTTGTAGTAGCTCTGC 3'), with an expected product size of 250 bp. To produce cDNA sequence templates, samples of 1 µg purified RNA were heated at 70°C for 5 min with 5 nmol oligo dT primers, and then

subjected to reverse transcription at 42°C for 1 h, in the presence of 2 µL dNTPs (10 mM), 4 µL 5X first strand buffer, 2 µL DTT (0.1 M), 200 U RNAse OUT and 40 U reverse transcriptase SuperScriptII (Invitrogen), in a total volume of 20 µL. The synthesis of cDNA was confirmed by the presence of housekeeping genes such as GAPDH (forward - 5' CCTCCTGCACCA CCAAC3' and reverse - 5' GCTGTGGGCAAGGTCATC 3') and NOTCH2 (forward - 5' TG TGGCCAACCAGTTCTCC 3' and reverse - 5' GGCAGTCATCAATATTCCTC 3') by PCR.

Primers were designed to anneal to two contiguous segments (potential exons) of the longest gapped sequence of each selected cluster. This way, the transcript amplification could be distinguished from genomic amplification, because of the presence of an intron in the latter. The program Primer3 (Rozen and Skaletsky, 2000) was used for primer design. PCR was then performed using 1 µL of a 1:3 dilution of single-stranded cDNA, 10 µM of each specific primer (Table 1), 2.5 µL PCR buffer (10X), 0.75 µL $MgCl_2$ (50 mM), 0.5 µL dNTP (10 mM), 0.25 µL of Taq Polymerase (5 U/µL), 5 µL betaine (5 M) (Henke et al., 1997) and 13 µL water. Reactions were performed with a basic cycle consisting of 30 s at 94°C, 45 s at optimal primer annealing temperature and 1 min at 72°C, for 35 cycles. Initial denaturation at 94°C for 4 min and final extension at 72°C for 7 min were carried out. The PCR products were resolved through an 8% polyacrylamide/TBE gel and DNA bands were silver-stained. The products that presented bands with expected sizes were submitted to electrophoresis through a 1% agarose/TBE gel stained with ethidium bromide. The bands were cut out and purified using the Concert™ Nucleic Acid Purification System (GibcoBRL), according to the manufacturer's instructions and then subjected to sequencing reaction by the use of the Big-Dye Terminator Kit (Applied Biosystems). The sequencing of the PCR product was carried out using an ABI 3100 automatic sequencer (Applied Biosystems). The chromatogram analysis was carried out using the Chromas software (Technelysium Pty Ltd.), and each sequencing FASTA file was aligned on genomic and transcriptome data using BLAT to confirm the specificity of the amplified products.

In silico expression analysis

Since the validated sequences showed alignment with predicted genes, we considered the predicted mRNA to search for virtual SAGE tags, taking 10-bp downstream to the more 3' *Nla*III restriction site CATG, and determined the tag frequency in a local SAGE databank containing 176 SAGE libraries. The number of normal and tumor-derived ESTs present within the genomic coordinates of these predicted mRNAs was also verified. To infer differential expression between normal and tumor tissue, a Bayesian statistics method was used for both SAGE and EST analysis (Vencio et al., 2004). Another way to foresee the importance of the expression of the predicted transcripts was by comparison of their cytogenetic localization with those described as RITE in the literature (Zhou et al., 2003), which indicates chromosomal regions associated with genes overexpressed in tumors.

RESULTS AND DISCUSSION

Of the 207,993 breasts and prostate ORESTES present in the HCGP database, 62,165 were no-match. A total of 62% of them (38,976) did not reach genome identity criteria, while 23,189 sequences mapped on the human genome were assembled into 14,908 clusters. Of these clusters, 976 containing full-length sequences and 1467 with ESTs from other projects were

removed for lack of novelty. We removed 1953 clusters comprising ESTs derived from single libraries and 8286 singletons without gaps upon their genomic alignment, totaling 12,682 clusters removed. The remaining no-match 2226 clusters were considered potential candidates. The average cluster size of the latter was 4.4 and 1/3 of them had a positive ESTSCAN score. These 2226 clusters could be divided into two sets: 187 spliced clusters and 2039 clusters without splicing. A total of 795 of 2.039 (39%) clusters without splicing but originating from more than two libraries had a positive ESTSCAN score. Splicing canonical sites were searched in the internal extremities of aligned EST segments, resulting in the selection of 94 spliced clusters, 65 of which were singletons (positive ESTSCAN score for almost 40%).

To further explore the potential of these clusters, we selected the clusters with the most stringent criteria to be considered genes for manual inspection. The 94 clusters selected were aligned with the updated human genome and transcriptome data, using BLAST and BLAT alignment search tools, and were manually checked by means of two graphic interfaces, the NCBI Map Viewer (http://www.ncbi.nlm.nih.gov/) and the UCSC (University of California Santa Cruz) Genome Browser (http://genome.ucsc.edu/). A total of 49.6% of the manually inspected clusters contained new sequences. The hits were divided into categories and subcategories to obtain a profile of the selected sequences (Table 1). A total of 71.9% of the selected sequences were localized in intragenic regions of known genes (RefSeq reviewed, provisional, validated, model, predicted, and *ab initio* predicted) either in introns or exons. The intronic, spliced expanded exon overlapping intron and spliced exonic-intronic hits, which corresponded to 42.2%, are probable low-expression alternative forms of the characterized genes. They may also come from overlapping genes in the sense or anti-sense orientation, and some, principally those aligned just with introns, may be non-coding RNAs (Kapranov et al., 2002). For the intronic fraction without splicing (8.5%), genomic and immature mRNA contamination is conceivable as well. A total of 7.4% of the selected sequences were in intergenic regions without annotation, 3.2% being spliced unpredicted genes.

Comparison was made of the 94 pipeline-selected clusters with 94 random clusters taken out from different steps of the pipeline application. These random groups were clusters with just one library and the singletons without gaps. This comparison showed us that the group of pipeline-selected clusters have more exonic hits (40.3 *vs* 1.4%, respectively), both in known and predicted genes, and less intronic and intergenic alignments (50.7 *vs* 98.6%), which suggests a more genomic and immature mRNA contamination rate for the last groups.

We selected 3 sequences from the prostate and 5 from the breast datasets for experimental validation. Experimental validation by PCR showed bands of expected size for five of them, two from prostate and three from breast datasets, which were then purified and sequenced. All five sequences showed the exact alignment, confirming that these no-match sequences in fact derive from human transcripts. With the exception of the mRNA correlate with 363833_OR sequence, the amount of ESTs present in the predicted gene regions (on average 180 ESTs) did not corroborate the expectation that these transcripts would be less abundant or difficult to detect when compared with the EST representation average of known genes (~136 EST/cluster from the full-length coverage). However, it seems that part of our sequences are still no-match not because of the low-EST abundance from their transcripts, but due to the fact that most of the ESTs do not overlap with the ORESTES, filling other gene portions, principally at the extremities. This suggests that a great part of the no-match dataset may be singletons or make up small clusters (3.6 EST/cluster) when compared with the average size of clusters

Table 1. Overall results of hits, classified according their categories and subcategories.

	Type of genomic hit	Number of hits	Percentage of original no-match sequences (%)	Observations
Intragenic (RefSeq Status: reviewed, provisional or validated)	Intronic without splicing	5	5.3	Probable alternative splice forms/overlapping genes/ncRNA and genomic contamination
	Intronic with splicing	16	17	
	Expanded exon with splicing overlapping intron	1	1	
	Spliced exon-intron	1	1	
	Exonic with splicing	8	8.5	Genes recently identified
	Exonic without splicing	0	0	
Intragenic (predicted gene - RefSeq status: model or predicted)	Intronic without splicing	3	3.2	Probable alternative splice forms of predicted genes/ overlapping genes/ncRNA
	Intronic with splicing	4	4.2	
	Expanded exon with splicing overlapping intron	0	0	
	Spliced exon-intron	5	5.3	
	Exonic with splicing	16	17	Corroboration of gene prediction
	Exonic without splicing	1	1	
ab initio prediction	Intronic with splicing	3	3.2	Corroboration of gene prediction/unpredicted exons
	Expanded exon with splicing overlapping intron	1	1	
	Spliced exon-intron	1	1	
	Exonic with splicing	3	3.2	
Intergenic (region without annotation)	With splicing	3	3.2	Genomic contamination/unpredicted genes/ncRNA
	Without splicing	4	4.2	
Alternative beginning or end	With splicing	0	0	Probable alternative splice form
	Without splicing	0	0	
Hits on opposite strands	-	17	18	Chimera
Several local and global disordered hits	-	2	2.1	Chimera
Total	-	94	100	-

ncRNA: non-coding RNA.

performed by ESTs of various projects (5.9 ESTs/cluster; Sogayar et al., 2004), probably because these ORESTES represent regions difficult to access using other methods. Nonetheless, the exons represented by some of them may belong to alternative transcripts expressed at very low levels, which would make them rare exons (Sakabe et al., 2003). In any case, we believe that this dataset should help complete information on several genes and their transcript variants. As to the 363833_OR predicted gene region, only 4 ESTs were found, which indicates a transcript poorly represented indeed.

Since a large portion of ORESTES originated from tumor tissue, there was also the possibility of some of the validated sequences being new transcripts involved in tumorigenesis. For the purpose of obtaining some information on differential expression between normal and tumor tissues, an *in silico* analysis was carried out. The expression of four predicted mRNAs (NT_011387.258, NT_006576.797, NT_005535.5, and NT_022184.1316), which we assumed corresponded to 363833_OR, AW996738, AW999051, and BE062313 ORESTES, respectively, were evaluated by SAGE using the most 3' tags of the predicted mRNAs. For the BF371440 sequence, the tag was extracted from a 3' poly-A tailed mRNA that covers the 3' end of its respective predicted mRNA NT_008183.629. The GenBank accession number of this mRNA is BC004287. The tag frequency from this transcript was found 435 times in the SAGE databank, distributed in 108 libraries, 82 from tumor and 26 from normal libraries. The tag from the AW996738 predicted mRNA, was found 31 times (17 libraries), 27 times (13 libraries) from cancer tissues and, of these, 13 (6 libraries) from breast tumor. The tag of the other predicted mRNA, corresponding to the AW999051 sequence, was found 24 times in 18 libraries, 18 times (15 libraries) from tumor tissues. The tag frequency was very low for the two other transcripts. An evaluation of normal and tumor ESTs present in the genomic region delimited by these predicted mRNAs was also performed. Bayesian statistics was performed in both cases, EST content and SAGE analysis (Vencio et al., 2004), in which we adopted the Bays error rate of less than 0.3 to consider that a transcript is differentially expressed, which means that the overlap area between the distributions of the two populations should be less than 30% (www.lbc.ludwig.org.br/sage/betabin). A statistically significant difference was found between tumor and normal SAGE tags to AW996738 (E = 0.2) and to BF371440 (E = 0.2). For AW999051, no statistically significant difference was found between tumors and normal tissues (E = 0.45). The scarce results for the 363833_OR and BE062313 transcript-derived tags resulted in a powerless statistical evaluation. The EST content showed no statistically significant difference between the amount of tumor and normal EST origin.

Furthermore, we looked for the cytogenetic localization of these transcripts to determine whether they are situated in some RITE (chromosomal regions of increased tumor expression; Zhou et al., 2003), which can indicate a potential overexpression in tumors, and, thus, warrant further attention. Two sequences, BF371440 and AW999051, were located in regions that demonstrate increased tumor expression in the same tissues as those from which they originated, prostate and breast, respectively. BE062313 was located in a RITE corresponding to brain, liver and pancreas. The other two validated sequences were not localized in RITE. We conclude that some of them, above all the BF371440 respective transcript, seem to be over represented in tumor tissues and could be used as molecular markers.

In summarizing our data mining for these 94 clusters, we conclude that although 20% were chimeras and 30% were known genes, 50% still are new genes or isoforms and that in this dataset it was possible to find new cancer molecular markers. Since the percentage of se-

quences with a positive ESTSCAN score in these 94 explored clusters was almost the same as in clusters without splicing and with splicing but without canonical sites, 40 and 39%, respectively, we can assume that in the 2226 selected clusters we will have 50% new sequences, including 939 new exons and 165 unpredicted genes. In the worst-case scenario, if we take into account our experimental validation rate (5/8), we still have about 690 new sequences revealing how no-match ORESTES datasets can contribute to the discovery of new genes and new isoforms and to define the prostate and breast transcriptome.

ACKNOWLEDGMENTS

We thank Dr. Sandro de Souza and compbio-LICR group for providing know-how and databases, Dr. Celine Pompeia for her helpful discussions and comments, Arthur Penha, Daniel Furtado, Jane Kaiano, and Maria Cristina Rangel for technical assistance, Saul G. Jachieri for the opportunity to develop this project, and Fundação Antonio Prudente, CAPES and Ludwig Institute for financial support.

REFERENCES

Adams MD, Kelley JM, Gocayne JD and Dubnick M (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252: 1651-1656.

Adams MD, Dubnick M, Kerlavage AR, Moreno R et al. (1992). Sequence identification of 2,375 human brain genes. *Nature* 355: 632-634.

Boguski MS, Lowe TM and Tolstoshev CM (1993). dbEST-database for "expressed Tags". *Nat. Genet.* 4: 332-333.

Bortoluzzi S, d'Alessi F and Danieli GA (2000). A novel resource for the study of genes expressed in the adult human retina. *Invest. Ophthalmol. Vis. Sci.* 41: 3305-3308.

Brentani H, Caballero OL, Camargo AA, da Silva AM et al. (2003). The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. *Proc. Natl. Acad. Sci. USA* 100: 13418-13423.

Burset M, Seledtsov IA and Solovyev VV (2000). Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.* 28: 4364-4375.

Camargo AA, Samaia HP, Dias-Neto E, Simao DF et al. (2001). The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. *Proc. Natl. Acad. Sci. USA* 98: 12103-12108.

Claverie JM (1997). Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.* 6: 1735-1744.

de Souza SJ, Camargo AA, Briones MR, Costa FF et al. (2000). Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. *Proc. Natl. Acad. Sci. USA* 97: 12690-12693.

Dias Neto E, Correa RG, Verjovski-Almeida S, Briones MR et al. (2000). Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proc. Natl. Acad. Sci. USA* 97: 3491-3496.

Ferreira EN, Pires LC, Parmigiani RB, Bettoni F et al. (2004). Identification and complete sequencing of novel human transcripts through the use of mouse orthologs and testis cDNA sequences. *Genet. Mol. Res.* 3: 493-511.

Galante PA, Sakabe NJ, Kirschbaum-Slager N and de Souza SJ (2004). Detection and evaluation of intron retention events in the human transcriptome. *RNA* 10: 757-765.

Henke W, Herdel K, Jung K, Schnorr D et al. (1997). Betaine improves the PCR amplification of GC-rich DNA sequences. *Nucleic Acids Res.* 25: 3957-3958.

Hudson TJ, Stein LD, Gerety SS, Ma J et al. (1995). An STS-based map of the human genome. *Science* 270: 1945-1954.

Kapranov P, Cawley SE, Drenkow J, Bekiranov S et al. (2002). Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 296: 916-919.

Kent WJ, Sugnet CW, Furey TS, Roskin KM et al. (2002). The Human Genome Browser at UCSC. *Genome Res.* 12: 996-1006.

Lander ES, Linton LM, Birren B, Nusbaum C et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.
Liang F, Holt I, Pertea G, Karamycheva S et al. (2000). An optimized protocol for analysis of EST sequences. *Nucleic Acids Res.* 28: 3657-3665.
Modrek B, Resch A, Grasso C and Lee C (2001). Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.* 29: 2850-2859.
Okubo K, Hori N, Matoba R, Niiyama T et al. (1992). Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nat. Genet.* 2: 173-179.
Rozen S and Skaletsky H (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* 132: 365-386.
Sakabe NJ, de Souza JE, Galante PA, de Oliveira PS et al. (2003). ORESTES are enriched in rare exon usage variants affecting the encoded proteins. *C.R. Biol.* 326: 979-985.
Schuler GD, Boguski MS, Stewart EA, Stein LD et al. (1996). A gene map of the human genome. *Science* 274: 540-546.
Sogayar MC, Camargo AA, Bettoni F, Carraro DM et al. (2004). A transcript finishing initiative for closing gaps in the human transcriptome. *Genome Res.* 14: 1413-1423.
Sorek R and Safer HM (2003). A novel algorithm for computational identification of contaminated EST libraries. *Nucleic Acids Res.* 31: 1067-1074.
Vasmatzis G, Essand M, Brinkmann U, Lee B et al. (1998). Discovery of three genes specifically expressed in human prostate by expressed sequence tag database analysis. *Proc. Natl. Acad. Sci. USA* 95: 300-304.
Vencio RZ, Brentani H, Patrao DF and Pereira CA (2004). Bayesian model accounting for within-class biological variability in Serial Analysis of Gene Expression (SAGE). *BMC. Bioinformatics*, online: 1-13.
Wang Z, Lo HS, Yang H, Gere S et al. (2003). Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. *Cancer Res.* 63: 655-657.
Wolfsberg TG and Landsman D (1997). A comparison of expressed sequence tags (ESTs) to human genomic sequences. *Nucleic Acids Res.* 25: 1626-1632.
Yu Y, Zhang C, Zhou G, Wu S et al. (2001). Gene expression profiling in human fetal liver and identification of tissue- and developmental-stage-specific genes through compiled expression profiles and efficient cloning of full-length cDNAs. *Genome Res*. 11: 1392-1403.
Zhou Y, Luoh SM, Zhang Y, Watanabe C et al. (2003). Genome-wide identification of chromosomal regions of increased tumor expression by transcriptome analysis. *Cancer Res.* 63: 5781-5784.