Identification of human exons overexpressed in tumors through the use of genome and expressed sequence data

Natanja Kirschbaum-Slager, Raphael Bessa Parmigiani, Anamaria Aranha Camargo and Sandro José de Souza


You might find this additional info useful...

Supplemental material for this article can be found at:
http://physiolgenomics.physiology.org/content/suppl/2005/05/12/00237.2004.DC1.html

This article cites 38 articles, 19 of which can be accessed free at:
http://physiolgenomics.physiology.org/content/21/3/423.full.html#ref-list-1

This article has been cited by 5 other HighWire hosted articles

- Global enhancement of nuclear localization-dependent nuclear transport in transformed cells
  Henna V. Kuusisto, Kylie M. Wagstaff, Gualtiero Alvisi, Daniela M. Roth and David A. Jans
  [Abstract] [Full Text] [PDF]

- Mammary Gland Selective Excision of c-Jun Identifies Its Role in mRNA Splicing
  Sanjay Katyar, Xuanmao Jiao, Sankar Addya, Adam Ertel, Yolanda Covarrubias, Vanessa Rose, Mathew C. Casimiro, Jie Zhou, Michael P. Lisanti, Talat Nasim, Paolo Fortina and Richard G. Pestell
  [Abstract] [Full Text] [PDF]

- Alternative splicing and genetic diversity: silencers are more frequently modified by SNVs associated with alternative exon/intron borders
  Jorge E. S. de Souza, Rodrigo F. Ramalho, Pedro A. F. Galante, Diogo Meyer and Sandro J. de Souza
  [Abstract] [Full Text] [PDF]

- Alternative splicing and genetic diversity: silencers are more frequently modified by SNVs associated with alternative exon/intron borders
  Jorge E. S. de Souza, Rodrigo F. Ramalho, Pedro A. F. Galante, Diogo Meyer and Sandro J. de Souza
  [Abstract] [Full Text] [PDF]

- Serial Analysis of Gene Expression in Adrenocortical Hyperplasia Caused by a Germline PRKARIA Mutation
  Anelia Horvath, Ludmila Mathyakina, Queenie Vong, Vanessa Baxendale, Alan L. Y. Pang, Wai-Yee Chan and Constantine A. Stratakis
  JCEM, February 1, 2012; 91 (2): 584-596.
  [Abstract] [Full Text] [PDF]

Updated information and services including high resolution figures, can be found at:
http://physiolgenomics.physiology.org/content/21/3/423.full.html

Additional material and information about Physiological Genomics can be found at:
http://www.the-aps.org/publications/pg
Physiological Genomics publishes results of a wide variety of studies from human and from informative model systems with techniques linking genes and pathways to physiology, from prokaryotes to eukaryotes. It is published quarterly in January, April, July, and October by the American Physiological Society, 9650 Rockville Pike, Bethesda MD 20814-3991. Copyright © 2005 by the American Physiological Society. ISSN: 1094-8341, ESSN: 1531-2267. Visit our website at http://www.the-aps.org/.

This information is current as of May 30, 2012.
Identification of human exons overexpressed in tumors through the use of genome and expressed sequence data

Natanja Kirschbaum-Slager, Raphael Bessa Parmigiani, Anamaria Aranha Camargo, and Sandro José de Souza.

Ludwig Institute for Cancer Research, São Paulo Branch, São Paulo, Brazil

Submitted 12 October 2004; accepted in final form 15 March 2005


Alternative splicing is one of the main sources of the variability found in the human transcriptome (3). There are four different types of alternative splicing: exon skipping/usage, alternative usage of a donor site, alternative usage of an acceptor site, and intron retention (20). Several bioinformatics analyses have indicated that at least one-half of all human genes undergo alternative splicing (7, 11, 17, 22, 23). In roughly 80% of these cases, alternative splicing invokes changes in the coding region (CDS) of genes, resulting in structural changes of the respective protein product (14, 23).

The biological impact of alternative splicing is perceptible, for example, in Drosophila, in which sex determination is triggered by alternative splicing of a master gene (25). Furthermore, ~15% of all human genetic diseases are believed to be caused by mutations in the splicing acceptor/donor sites, generating changes in the splicing pattern of one or more genes, which implies that alternative splicing also plays an important role in pathogenicity (19).

An apparent link between certain cancer types and alternative splicing is being investigated (for a review, see Caballero et al., Ref. 8). Several splicing variants from different genes, including cd44, wt1, cd79b, bin1, and Syk, have been shown to be associated with different aspects of tumorigenesis (1, 10, 13, 26, 32). The increasing amount of cDNA libraries constructed from a diversity of both tumor and normal tissues and cell lines allows several types of computational analyses. This, together with the release of the final sequence of the human genome (16, 31), permits genome-wide analyses of alternative splicing and the search for tumor-associated splicing variants. Several groups have performed such analyses and have reported the differential expression of splicing variants in tumors (15, 33–35). None of these studies, however, systematically verified the expression pattern of the prototype variant of the same candidate gene (33, 15). Hence, it cannot be ruled out that the variants selected by their analyses as being tumor specific are variants of genes that are generally overexpressed in tumors. Furthermore, none of those studies has investigated the expression of splicing variants within tumors of one specific tissue.

Here, by using strict selection and statistical criteria, we were able to screen the genome for exons overexpressed in tumors. Tumor-associated exons are those that appear preferentially in splicing isoforms found to be overexpressed in tumors. Such exons could be of major diagnostic value, allowing the early detection of tumors based on their specific expression. New epitopes encoded by tumor-associated exons may be targeted by antibodies as well. Eventually, this should permit drug design, as the protein encoded by a spliced variant may be a therapeutic target. Here, we show by experimental, statistical, and literature validation that our set of candidates is enriched with bona fide tumor-associated splicing variants.

Materials and Methods

cDNA mapping and clustering. All human cDNAs available in dbEST (July 2002, Ref. 4) and mRNA sequences from known human genes from Unigene release 153 (29) were aligned to the masked human genome sequence [build 29, obtained from the National Center for Biotechnology Information (NCBI)] by use of pp-Blast (27), an implementation of MEGABLAST (37) for a parallel cluster. The parameters used in MEGABLAST were: -f T -J F -F F -W 24. The MEGABLAST output was parsed, and a MySQL database was loaded with the mapping information. Spurious hits were excluded from the mapping database by use of an additional set of alignment criteria. These include a minimum degree of identity for a cDNA/gene alignment set to 93% or at least 45% of the total expressed sequence tag (EST) length or 55% of the total length of the full-insert sequence. Furthermore, for sequences mapping to more than one
Serial analysis of gene expression tag assignment. A virtual serial analysis of gene expression (SAGE) tag is a prediction of the 10-bp sequence downstream of the 3’-most NlaIII site of the transcript that might theoretically be produced by a SAGE experiment (5). One representative full-insert mRNA was selected from those candidate clusters that included at least one full-insert mRNA showing at least either a poly A signal and/or a poly A tail. This full insert was then assigned a virtual SAGE tag (5). The tag was assigned only to the 3’-most NlaIII site of the transcript. This tag was used to query all SAGE libraries of the same tissue in which we characterized the putative overexpressed exon. The frequency of each tag was counted in tumor and normal libraries of the same tissue.

Again a Z-statistic was calculated

\[ Z = \frac{(p_t - p_n)}{\sqrt{1 - p(1/n_t + 1/n_n)}} \]

For a given exon, \( p_t \) and \( p_n \) are the expression frequencies of the specific 3’-most SAGE tag in tumor and normal libraries, respectively, in a specific tissue (the no. of tumor or normal ESTs containing the specific exon \( \div \) total no. of tumor or normal ESTs from all libraries). To minimize sampling bias of small libraries, we only took into account libraries that had at least the size of the smallest library in which a transcript containing the specific exon was found in the specific tissue. The \( p \) is the geometric average frequency of the exon in tumor and normal libraries, and \( n_t \) and \( n_n \) are the numbers of ESTs in the normal and tumor libraries, respectively, taken into account for each specific exon in each tissue. (In each tissue, Z-values having a \( P \leq 0.05 \) were considered significant. (It should therefore be noted that the statistically significant candidates still have a probability \( P < 0.05 \) of being a false-positive candidate.)

RESULTS

Transcriptome database. The database used in this work contains data obtained from alignments of all cDNA sequences to the human genome sequence (12, 28). In addition to the representation of all data concerning the alignment and clustering of the sequences, the database also contains binary matrices that were constructed for each transcript (28). In such a matrix, a transcribed exon is represented by a one (1) and an absent exon is represented by a zero (0). This approach facilitates the analysis of exon skipping/exon usage throughout the genome and the comparison of the different transcripts and exons with each other.

Our database contains 3,475,514 expressed sequences from 7,167 cDNA libraries from different tissues (see Table 1), of which 52,903 represent full-insert sequences (completely sequenced cDNA clones). Four thousand, two hundred and forty-nine (4,249) of these libraries were constructed from tumor samples and tumor cell lines, generating 1,427,390 sequences, while the remaining 2,918 libraries were constructed from normal samples, generating 2,048,124 sequences. We will refer to libraries constructed from either tumor samples or tumor cell lines as tumor libraries. Our analysis was performed on both normalized and nonnormalized libraries.

Database validation. Our clustering strategy (28) generated 318,272 cDNA clusters, 21,306 containing at least one full-insert mRNA. Of all clusters containing at least one full-insert mRNA, 52% underwent exon skipping (12), which is agree-

\[ Z = \frac{(p_t - p_n)}{\sqrt{1 - p(1/n_t + 1/n_n)}} \]

For a given exon, \( p_t \) and \( p_n \) are the expression frequencies of the specific 3’-most SAGE tag in tumor and normal libraries, respectively, in a specific tissue (the no. of tumor or normal tags \( \div \) total no. of tumor or normal tags from all libraries in that tissue). The \( p \) is the geometric average frequency of the tag in tumor and normal libraries, and \( n_t \) and \( n_n \) are the numbers of tags in the normal and tumor libraries taken into account for each specific exon. Z-values having a \( P < 0.05 \) were considered significant (It should therefore be noted that the statistically significant candidates still have a probability \( P < 0.05 \) of being a false-positive candidate.)

Experimental validation. Total RNA derived from five different normal human tissues (lung, prostate, breast, brain, colon) was purchased from Clontech Laboratories and used for cDNA synthesis. Human tumor cell lines were obtained from the American Type Culture Collection (ATCC) and maintained in appropriate medium as recommended by this organization (http://www.atcc.org). The following human tumor cell lines were used: A172 and T98G (glioblastoma), DU145 and PC3 (prostate), MCF-7 and MDA-MB-231 (breast), H1155 and H358 (lung), and SW480 (colon).

Patient samples were obtained from the Hospital A. C. Camargo tumor collection and prepared by manual dissection. All patient samples were collected after explicit informed consent, and the study was approved by the Institutional Ethics Committee.

Total RNA was extracted from tumor cell lines and tumor/normal patient samples by a conventional CsCl-guanidine thiocyanate gradient method (9), and RNA integrity was analyzed using agarose gels. Genomic DNA contamination of the total RNA was tested with PCR, using hMLH1 primers located at intronic sequences flanking exon 12 (forward, 5’-TGG TGT CTC TAG TTC TGG-3’; reverse, 5’-CAT TGT TGT AGT AGC TCT GC-3’).

Reverse transcription was carried out using the Superscript First Strand Synthesis Kit, according to the manufacturer’s instructions (Invitrogen). RT-PCR reactions were carried out in a 25-μl reaction mixture containing 1 μl of cDNA, 1× Taq DNA polymerase buffer, 0.1 mM dNTPs, 6 pmol of each primer (for sequences of primers, see Supplemental Material; available at the Physiological Genomics web site), 1 mM MgCl₂, and 1 U Taq DNA polymerase (Invitrogen). Standard PCR conditions were as follows: 4 min at 94°C (initial denaturation), 35 cycles of 45 s at 94°C, 45 s at 58°C, and 1 min at 72°C, with a final extension step of 10 min at 72°C. RT-PCR products were analyzed on 3% silver-stained polyacrylamide gels and on 2% ethidium bromide-agarose gels. Sequencing reactions were carried out using DYEnamic (ET Terminator Cycle Sequencing Kit, Amersham Pharmacia) and an ABI 377 prism sequencer (Perkin Elmer), according to the supplier’s recommendations.
ment with the splicing rate reported in the literature (11, 17, 23). Considering all clusters in the database, we found that 12,648 present at least one exon-skipping variant. Our analysis was performed on this latter set of clusters.

The suitability of our exon-skipping database for the current analysis was manually evaluated through the analysis of 61 genes described in the literature to have at least 2 splicing isoforms. Compared with the literature, 62% of those genes (38/61) were shown to have the same or a larger number of variants in the exon-skipping database than the number of variants published (see Supplemental Table S1). It should be noted that, although examples from the literature include all types of alternative splicing, our database considers only exon skipping/usage. This validation step confirmed that the exon-skipping database sufficiently covers the repertoire of splicing variants represented in the sequence databases.

Tumor-specific exons. We screened our database for potential tumor-associated exons, which appear in isoforms found to be exclusively expressed in tumor samples and tumor cell lines. Our clustering strategy and matrix representation allowed us to screen our database for exons that were not expressed in transcripts from any normal tissue but were expressed in transcripts from tumor libraries. For each gene, at least one transcript showing exon skipping was chosen to represent the cluster; the exon-skipping event should be confirmed by at least two cDNA sequences from different libraries. We screened for variants that would show the expression of an exon at the exact position of the skipped exon in this prototype transcript. Exons fitting into this category had to be flanked by at least two other exons; they should be represented by sequences derived from tumor libraries only. We increased the stringency of our analysis by only selecting those exon usage events that were confirmed by at least two cDNA sequences from different libraries. We screened for variants that would show the expression of an exon at the exact position of the skipped exon in this prototype transcript. Exons fitting into this category had to be flanked by at least two other exons; they should be represented by sequences derived from tumor libraries only. We increased the stringency of our analysis by only selecting those exon usage events that were confirmed by at least two cDNA sequences from different libraries. Because of these stringent criteria, we were able to identify only 11 tumor-specific exons from 11 different genes (see Supplemental Material). The variants skipping these exons were expressed in several normal tissues.

Tumor-associated exons expressed in specific tissues. On the basis of the low number of candidates identified by our first approach, we decided to investigate whether a given exon could be tumor specific when its expression pattern was analyzed within one specific tissue. The search criteria for our candidates and the number of clusters filtered in each step are summarized in Fig. 1. A certain variant was defined as a candidate when it was associated with tumor samples and cell lines within one tissue only, although it could appear in normal samples of other tissues. For this purpose, all libraries were divided into tissue groups according to their annotations. Within each of the 60 selected groups, libraries were subdi-

---

**Table 1. No. of tumor and normal libraries, ESTs, and tissue groups in the exon-skipping database**

<table>
<thead>
<tr>
<th>Libraries</th>
<th>Tissue Groups</th>
<th>ESTs</th>
</tr>
</thead>
<tbody>
<tr>
<td>All libraries</td>
<td>7,167</td>
<td>60</td>
</tr>
<tr>
<td>Tumor libraries</td>
<td>4,249</td>
<td>42</td>
</tr>
<tr>
<td>Normal libraries</td>
<td>2,918</td>
<td>53</td>
</tr>
</tbody>
</table>

Both types of libraries include cell lines and patient tissue. EST, expressed sequence tag.

---

![Flow chart](https://via.placeholder.com/150)
vided into those derived from either tumor or normal tissue (see Supplemental Table S2). Only those 37 groups that included both tumor and normal libraries were used. This tissue-specific analysis increased the number of candidates to 2,271 genes, including 4,916 tumor-associated exons within different tissue types (a list of all candidate genes is available; see Supplemental Table S3). Of these genes, 2,108 contained at least one full-insert mRNA (containing 4,647 candidate exons).

Statistical filter for the tumor-associated variants. Tumor association of each of the candidate exons was tested for its statistical significance. A $Z$-score was calculated for each candidate exon (see MATERIALS AND METHODS) per tissue (33). This statistical approach takes into account the total number of ESTs for each tissue group in either normal or tumor libraries (see MATERIALS AND METHODS). Of the total number of candidate exons (4,916), 2,878 (59%) were shown to be significantly associated with tumors ($P < 0.05$). Of all candidates potentially associated with brain tumors (269 candidate exons), 233 (87%) exons presented a significant $Z$-score ($P < 0.05$). For prostate, lung, breast, and colon, 192 of 203 (95%), 193 of 239 (81%), 272 of 461 (59%), and 266 of 847 (31%) candidate exons, respectively, were shown to be significantly associated with tumors within the respective tissue (Table 2, column 3, and Supplemental Table S4).

Experimental validation. Seven candidates were randomly selected to be screened for expression of their putative tumor-associated exons by RT-PCR on cDNA from normal tissues and tumor cell lines. Five candidates were selected from brain, one from breast, and one from prostate, all of them passing the statistical test ($P < 0.05$) in the respective tissue. Three primers were designed for each candidate: two on the exons flanking the candidate tumor-associated exon (flanking primers), and one on the exon itself (specific primer). The products from the reaction using one flanking and one specific primer showed overexpression of the candidate variant in the respective tumor cell line (Fig. 2, data for 3 candidates). However, when using the two flanking primers, we observed that the variant skipping the exon was also overexpressed in the tumor cell lines (Fig. 2). This raised the possibility that our analysis was inflated with genes overexpressed in tumors instead of tumor-associated variants.

SAGE analysis. On the basis of the above observations, we implemented an additional filter selecting those candidate exons that did not belong to genes overexpressed in tumors. A virtual SAGE analysis was performed to verify whether the candidate genes were overexpressed in tumors from the respective tissue (5). We computationally assigned a virtual SAGE tag to one full-insert transcript of each gene (see MATERIALS AND METHODS) and statistically verified the tumor-to-normal ratio for

<table>
<thead>
<tr>
<th>Table 2. No. of candidate exons after original screening criteria, after statistical filter, and after SAGE filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>After Initial Criteria</td>
</tr>
<tr>
<td>------------------------</td>
</tr>
<tr>
<td>Tumor-associated exons in all tissues</td>
</tr>
<tr>
<td>Tumor-associated exons in brain</td>
</tr>
<tr>
<td>Tumor-associated exons in breast</td>
</tr>
<tr>
<td>Tumor-associated exons in prostate</td>
</tr>
<tr>
<td>Tumor-associated exons in lung</td>
</tr>
<tr>
<td>Tumor-associated exons in colon</td>
</tr>
</tbody>
</table>

SAGE, serial analysis of gene expression.

**Fig. 2.** Experimental validation of 3 brain tumor-associated candidates after the statistical filter. Candidates that passed the statistical filter were randomly chosen to be screened for expression of their tumor-associated exons by RT-PCR on cDNA from normal tissues and tumor cell lines. Three primers were designed for each candidate: 2 on the exons flanking the candidate tumor-associated exon and 1 annealing to the tumor-associated exon itself. We performed 2 sets of reactions, 1 using the flanking primers, which should amplify both variants, and 1 using 1 flanking and the specific primer, which should amplify only the variant expressing the candidate exon. N1, normal whole brain tissue; T1, T98G glioblastoma cell line; T2, A172 glioblastoma cell line; No, “no DNA” control. Either flanking primers or specific primers were used for the amplification of variants of the following genes: THC211630 (AJ010070; A), CDK-2 (NM_052827; B), and calponin 2-CNN2 (AK057960; C). The products from the 2nd reaction (using the primers annealing to the exon itself) showed overexpression of the candidate variant in the respective tumor cell line. However, the variant skipping the specific exon was also overexpressed in the tumor cell lines. Amplification of GAPDH as a positive control is shown for all samples in D.

Physiol Genomics • VOL 21 • www.physiolgenomics.org
that respective tag. All candidate genes showing a statistically significantly higher tag count in tumors were excluded from the list of candidates (see MATERIALS AND METHODS).

After this additional filter, 638 candidate genes, including 1,386 exons, remained in our list of candidates (Table 2, column 4, and Supplemental Table S5). The distribution of transcripts with more than one candidate exon is shown in Supplemental Fig. S1. The final list contained 172 candidate genes containing potential tumor-associated exons in brain, 183 in breast, 138 in prostate, 156 in lung, and 235 in colon.

We selected a few candidates for experimental validation. Using RNA extracted from tumor cell lines and normal tissues, we observed that, of the 10 candidates with conclusive results, 4 candidate genes showed that the variant containing the candidate exon was overexpressed in either brain, lung, or colon tumor cell lines, whereas the exon-skipping prototype was not (Fig. 3). The other six candidates showed a pattern similar to those in Fig. 2: both the exon-skipping variant and the variant including the selected tumor-associated exon were overexpressed in tumor tissue (results not shown).

Three of the four positively validated cases in the cell lines were also validated in patient samples (Fig. 4). Interestingly, when testing two of the six cases that were negative in cell lines, both were positively validated in some of the patient samples (Fig. 5).

Five of the six exons validated in patient samples are located inside the coding region of their respective gene. Of those five, the length of four exons is not a multiple of three and can therefore be expected to cause a change in the reading frame of its gene.
To analyze the functional characteristics of the final list of genes, we compared our candidates to a list of 1,127 cancer-related (CR) genes (15a). This list was a manually curated compilation based on queries of various public databases using the words “cancer” and “tumor” (for more details, see Brentani et al., Ref. 15a). The CR genes in the list constitute 5.3% of the known genes of our transcriptome database. When analyzing our 638 candidates, we found an overlap of 60 candidates (9.4%) in the CR list. Among these genes, we found Syk and bin1, which are known to have tumor-associated variants (13, 33) (for a whole list, see Supplemental Table S6). Thus there is an excess of
cancer-related genes in our final list of candidates (chi-square = 7.97, 1 degree of freedom, \( P = 0.005 \)). Comparing these results to a simulation of 200 randomly chosen sets of 638 clusters out of all UniGene clusters, we found that none of these sets presented >60 CR genes (\( P < 0.005 \)).

The Gene Ontology (GO) terms of the final list of 638 candidates were obtained with the GOTM program (36). For 437 of the 638 genes, a GO term could be assigned (see Supplemental Fig. S3 for an overview of the distribution of the GO terms in the different GO categories and Supplemental Fig. S4 for all levels of the GO tree). The program GOstat (2) was used to analyze whether any GO category was overrepresented in our final list of candidates relative to the representation of all ontology terms in the ontology database (see Supplemental File S1). In each category, the lowest \( P \) value resulting in biologically meaningful GO terms was used. In the category “biological process,” using a stringent \( P \) value cutoff of \( 10^{-5} \), we found the GOs “intracellular protein transport” and “cell growth and maintenance” to be significantly overrepresented. In the category “molecular process,” the GOs “actin binding,” “receptor activity,” “cytoskeletal protein binding,” and “ATP binding” were significantly overrepresented (cutoff \( P \) value of 0.001). Finally, in the category “cellular components,” the GO “peroxisome” was significantly overrepresented (\( P \) value cutoff = 0.001).

**Literature validation.** In one reported study (33) of tumor-associated splicing variants, experimental validation was performed in 76 genes chosen either by statistical criteria or by knowledge of their tumor association. All 76 candidates were experimentally validated (M. P. Lee, personal communication).

To validate our candidates once more, we verified whether we could find any of our candidates in the published list of experimentally validated genes (Table 3). Thirteen of the validated candidates of this reported work were found in our initial set of candidates before the SAGE analysis. In our final list of candidates, we could only find three of their candidates.

**Table 3. Overlap of our candidates with experimentally validated candidates from Wang et al. (33)**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Overlap of Candidates Before SAGE Filter</th>
<th>Overlap of Candidates After SAGE Filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>NME1</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>CDC25C</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>DVL1</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>ERCC1</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>GSS</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>GTF3C1</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>IRAK1</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>NIKTR</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>POLB</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>RAD51</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>SHC1</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>ST5</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>TNFRSF1A</td>
<td>+</td>
<td>–</td>
</tr>
</tbody>
</table>

**Fig. 5.** We evaluated the expression pattern of 2 more candidates that were originally negative in the validation using tumor cell lines. A: the gene NADH dehydrogenase (ubiquinone) Fe-S protein 2, 49 kDa (NADH-coenzyme Q reductase) (BC001456), was validated in paired lung samples. N1–N4, and T1–T4, same paired lung normal/tumor patient samples, respectively, as in Fig. 4C; No, no DNA control. The products from the reaction using the primers annealing to the exon itself showed tumor overexpression of the exon in 3 of 4 paired samples. The primers flanking the candidate exon show that the variant expressing the candidate exon was overexpressed in 3 tumor samples, whereas the prototype was not overexpressed in tumors. B: validation of the gene “proteasome (prosome, macropain) 26S subunit, non-ATPase, 10 (NM_002814)” in prostate normal pool and 3 patient tumor samples. The products from the reaction using the primers annealing to the exon itself showed tumor overexpression of the exon in 3 patient tumor samples. The primers flanking the candidate exon show that only the variant including the candidate exon is overexpressed in patient tumor samples. NP, normal pooled prostate cDNA; T1–T3, different prostate tumor patient samples; No, no DNA control.
The same comparison with a different study (35) indicated an overlapping of 21 candidates of 89 published genes (Table 4). Interestingly, we observed that 11 of the candidate genes in this last report (35) presented an overexpression in tumors as evaluated by SAGE. Taken together, these comparisons highlight the importance of filtering off genes generally overexpressed in tumors to increase the likelihood of finding bona fide tumor-associated splicing variants.

**DISCUSSION**

The characterization of splicing variants associated with tumors is critical for the development of new diagnostic and therapeutic strategies for the treatment of cancer. Few attempts have been made to search the human genome for tumor-associated splicing isoforms (35, 33, 15). An interesting aspect of these studies is the fact that none of them take into consideration whether the differential expression was specific to the respective splicing variant or common to all transcripts from that gene. Although some of these reports provided statistical arguments corroborating the association between the splicing isoform and tumors, most of them lack experimental validation. Wang et al. (33) showed experimental validation for the gene RABIA (Fig. 2 in Wang et al., Ref. 33). There, it is possible to see that, in some samples, the prototype variant is also overexpressed in tumors. In our attempt to define exons overexpressed in tumors, we faced the same problem.

The clustering of all human cDNAs onto the human genome sequence allowed us to focus our strategy on determining the expression pattern of all exons in the human genome. We were able to seek for exons that were exclusively represented by transcript sequences derived from tumor cDNA libraries. A broad computational analysis revealed that only 11 exons were found to be expressed in tumors with no expression at all in normal tissues. This motivated us to search for exons expressed only in tumor tissues or tumor cell lines within a specific tissue. Using this strategy, we found 4,916 tumor-associated exons.

The expression pattern of all variants was defined based on the annotation provided with the cDNA libraries publicly available. This information, however, is not always precise and clear. Because further tissue characterization of a transcript is dependent on this information, efforts are currently being made to verify the real character of all publicly available libraries (18).

To verify whether the presence of data from normalized libraries would compromise any analyses based on the relative expression levels of transcripts, we tested whether our set of candidates was inflated with sequences derived from normalized libraries. We found a number proportional to the total number of normalized data in our initial set (30% of candidates, 30% of normalized data in the initial set). This excludes the possibility that our set of candidates is enriched with artifacts due to data from normalized libraries.

To refine our analysis of tumor-associated exons, the candidates were further screened by a statistical analysis of each exon and, for a few cases, by RT-PCR validation in tumor cell lines. Roughly 41% of our candidate exons were excluded by the statistical filter.

Nowadays, experimental analysis by RT-PCR is one of the most specific ways of verifying the expression pattern of mRNA transcripts. However, while amplifying two different variants by the use of two flanking primers, competition in transcript amplification may not reflect correctly the intrinsic difference in the expression level of the two transcripts. When using a specific primer for one exon only, however, the primers may be of such specificity that they may amplify a maximum of the transcript regardless of its expression level within the cell. More sensitive methods like real-time PCR or single molecule profiling may be used to better quantify splicing variants (30, 38).

Experimental validation also showed that the whole gene, not only the candidate exon, was overexpressed in tumor cell lines. We found support for this when we performed a SAGE analysis for all of our candidates. For this approach, we assumed that the 3’-most SAGE tag is representative of the most abundant transcript of the candidate genes. We also assumed that the prototype is more abundantly expressed than the candidate variant and that the SAGE tag count is therefore an indication of the prototype expression pattern. We found that ~52% of our candidates obtained after the statistical filter represented genes overexpressed in tumors. All those cases were excluded, and a new list of candidates was produced containing 1,386 exons. Validation with those candidates showed a success rate of 40% (4/10) when tumor cell lines were used. When we used a panel of patient samples, the success rate was much higher, ~85% (5/6). This probably occurs because of both the limitations of RT-PCR and the still-limited number of SAGE libraries available today. Furthermore, the heterogeneity of tumor samples and cell line cultures provides another variable to the whole system. Only a large-scale validation scheme will allow the definitive test of our bioinformatics pipeline. However, here we show that the combination of our computational analysis with experimental validation is successful in screening for real cancer-associated exons at a success rate that would not be achieved using either a computational analysis or experimental validation alone.

Our final list of candidate genes is enriched with cancer-related genes (P = 0.005). As stated before, it is likely that

**Table 4. Overlap of our candidates with those of Xu and Lee (35), having log score >3**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Overlap of Candidates Before SAGE Filter</th>
<th>Overlap of Candidates After SAGE Filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>BZW2</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>CCT3</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>CGI-31</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>F11R</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>HGD</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>HLA-DMB</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>HNRPF</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>HNRPK</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>KIF2C</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>MBP</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>MED8</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>MGC11257</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>MORF4L2</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>NGLY1</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>NUDE1</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>SAD1</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>SMUG1</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>SPC18</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>TPM1</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>WDR4</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>WRB</td>
<td>+</td>
<td>–</td>
</tr>
</tbody>
</table>
variants associated with cancer are found in genes that are related to cancer. On the other hand, this does not mean that variants from genes not involved in cancer would not have a functional impact on tumorigenesis (35). An ontology analysis also suggested that genes involved in intracellular protein transport and cell growth and maintenance are overrepresented in our final list of candidates. One could expect that any change in the expression level of splicing variants of genes that are connected to cell cycle and maintenance might influence cell transformation. In the category cellular components, the GO peroxisome was significantly overrepresented. Interestingly, each of the peroxisome proliferator-activated receptor (PPAR) isotypes, for example, has been shown to be involved in the pathogenesis of several tumors. PPARα induces hepatocarcinomas; PPARγ has an anti-proliferation, pro-apoptotic effect and is therefore thought to have an anti-carcinogenic effect; and PPARβ/δ is involved in the control of cell proliferation and apoptosis (21). Further investigation on the impact of the overexpression of the variants of any genes involved in the above pathways might give some insight on the possible function of specific splicing variants.

To our knowledge, this is the first report that attempts to search specifically for exons overexpressed in tumors while excluding genes that are generally overexpressed in the same tumors. Such exons may provide valuable information for future investigations on the regulation of tumor-associated alternative splicing. Finally, further experimental analysis will validate the extent to which our candidate exons are of potential diagnostic and/or therapeutic value.

ACKNOWLEDGMENTS

We thank Dr. Ricardo R. Brentani, Dr. Helena P. Brentani, Maria D. Vibranovski, Noboru J. Sakabe, and Pedro A. F. Galante for helpful discussions and/or careful reading of the manuscript.

GRANTS

N. Kirschbaum-Slagter is supported by PhD fellowships from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) (to 12/03) and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) (from 01/04). R. B. Parmigiani is supported by a PhD fellowship from CAPES. This project was supported by a Centros de Pesquisa, Inovação e Difusão Grant from FAPESP.

REFERENCES


