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

Physiol. Genomics 21:423-432, 2005. First published 22 March 2005;
doi:10.1152/physiolgenomics.00237.2004

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/.

Physiol Genomics 21: 423–432, 2005.
First published March 22, 2005; doi:10.1152/physiolgenomics.00237.2004.

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, Sao Paulo, Brazil

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

Slager-Kirschbaum, Natanja, Raphael Bessa Parmigiani, Anamaria Aranha Camargo, and Sandro José de Souza. Identification of human exons overexpressed in tumors through the use of genome and expressed sequence data. *Physiol Genomics* 21: 423–432, 2005. First published March 22, 2005; doi:10.1152/physiolgenomics.00237.2004.—Alternative splicing is one of the major sources of the large transcriptional diversity found in human cells. Splicing variants have been shown to be associated with features like spreading and progression in several human tumors. Therefore, such variants may be of great importance as both diagnostic and therapeutic tools. Here, by using a set of criteria regarding the expression pattern of splicing variants and statistical analyses, we were able to screen the genome for exons overexpressed in tumors of specific tissues. However, as in other analyses attempting to identify tumor-associated variants, our list of candidates was seriously inflated with cases of genes differentially expressed in tumors. To exclude these cases and increase the probability of finding bona fide regulated splicing variants, we performed a serial analysis of gene expression (SAGE), excluding those genes that were shown to be upregulated in tumors. This allowed us to predict the overexpression of single exons in specific tumors. Our final group of candidates includes 1,386 exons belonging to 638 genes. Experimental validation of a few candidates in normal tissue, tumor cell lines, and patient samples suggests that most of these candidates are indeed tumor-associated exons. Further functional classification of our candidate genes shows that our final list is slightly inflated with cancer-related genes.

alternative splicing; tumor; transcriptome; serial analysis of gene expression

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/genome alignment set to 93% over 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

Article published online before print. See web site for date of publication (http://physiolgenomics.physiology.org).

Address for reprint requests and other correspondence: S. J. de Souza, Ludwig Institute for Cancer Research, São Paulo Branch, Rua Prof. Antonio Prudente 109, 4 andar, São Paulo, 01509-010, SP, Brazil (e-mail: sandro@compbio.ludwig.org.br).

location on the genome, a score associated with a higher identity over a longer alignment was assigned. Clustering of cDNA sequences was based on their genomic coordinates as described by Sakabe et al. (28). Briefly, if two sequences shared at least partially the same gene structure, they were joined into the same cluster. If no exon/intron boundary was defined, a sequence had to have at least a 100-bp overlap with another sequence at the genome level to be added to the respective cluster.

Construction of the binary matrices. All sequences were represented as binary matrices, and each expressed exon was represented by 1 (one) and each skipped exon by 0 (zero). Variants were defined to skip an exon when they included two flanking exons next to an absent one (represented as 10+1, meaning that at least one exon is skipped between two flanking exons).

Z-statistics. After a screening for variants that included exons at the exact position of an exon skipping in another variant of the same cluster, a *Z*-statistic was calculated for each exon. This way, the probability of tumor association of the exon to a specific tissue, based on the numbers of ESTs confirming the variant in either tumor or normal tissue, was evaluated (33)

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

For a given exon, p_t and p_n are the expression frequencies of the exon in tumor and normal tissues, respectively, in a specific tissue (the no. of tumor or normal ESTs containing the specific exon ÷ 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_n and n_t 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.)

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 *Nla*III 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

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

For a given gene, 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 ÷ 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_n and n_t are the numbers of tags in the normal and tumor libraries taken into account for each specific exon. *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.)

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 appropriated 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$^-$ (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] 1 mM $MgCl_2$, 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 8% 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.

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% undergo exon skipping (12), which is in agree-

[1] The Supplemental Material for this article (Supplemental Tables S1–S6, Supplemental Figs. S1–S4, Supplemental File S1) is available online at **http://physiolgenomics.physiology.org/cgi/content/full/00237.2004/DC1**.

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

	Libraries	Tissue Groups	ESTs
All libraries	7,167	60	3,475,514
Tumor libraries	4,249	42	1,427,390
Normal libraries	2,918	53	2,048,124

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

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 derived from different tumor 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-

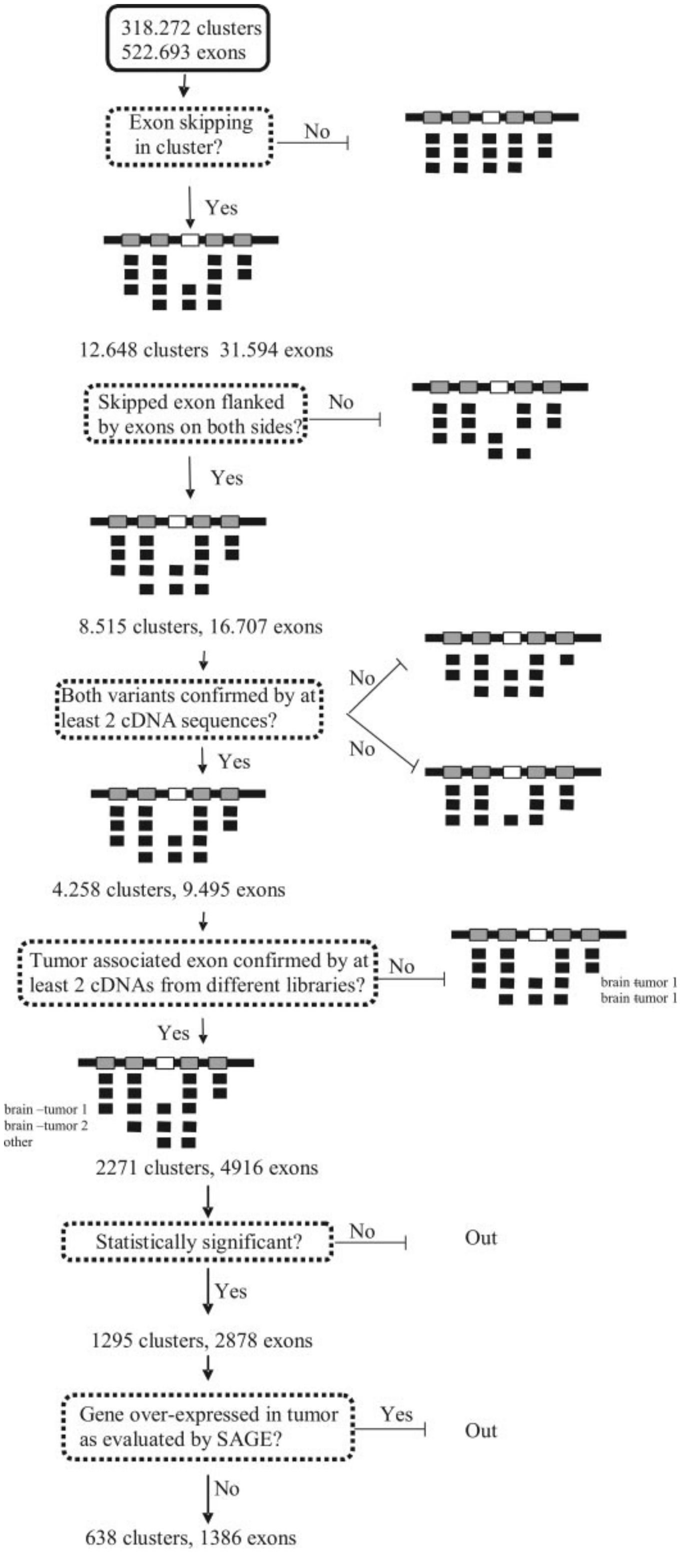
318.272 clusters
522.693 exons
Exon skipping in cluster?
No
Yes
12.648 clusters 31.594 exons
Skipped exon flanked by exons on both sides?
No
Yes
8.515 clusters, 16.707 exons
Both variants confirmed by at least 2 cDNA sequences?
No
No
Yes
4.258 clusters, 9.495 exons
Tumor associated exon confirmed by at least 2 cDNAs from different libraries?
No
brain tumor 1
brain tumor 1
Yes
brain –tumor 1
brain –tumor 2
other
2271 clusters, 4916 exons
Statistically significant?
No
Out
Yes
1295 clusters, 2878 exons
Gene over-expressed in tumor as evaluated by SAGE?
Yes
Out
No
638 clusters, 1386 exons

Fig. 1. Flow chart describing the approach used here to identify tumor-associated exons. Thick black lines with white boxes represent clusters; the black boxes under the clusters represent the exons present in the cDNA sequences that align to the cluster. The nos. of genes and exons obtained after each step of the screening are listed.

Table 2. *No. of candidate exons after original screening criteria, after statistical filter, and after SAGE filter*

	After Initial Criteria	After Statistical Filter	After SAGE Filter
Tumor-associated exons in all tissues	4,916	2,878	1,386
Tumor-associated exons in brain	269	233	172
Tumor-associated exons in breast	461	272	183
Tumor-associated exons in prostate	203	192	138
Tumor-associated exons in lung	239	193	156
Tumor-associated exons in colon	847	266	235

SAGE, serial analysis of gene expression.

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

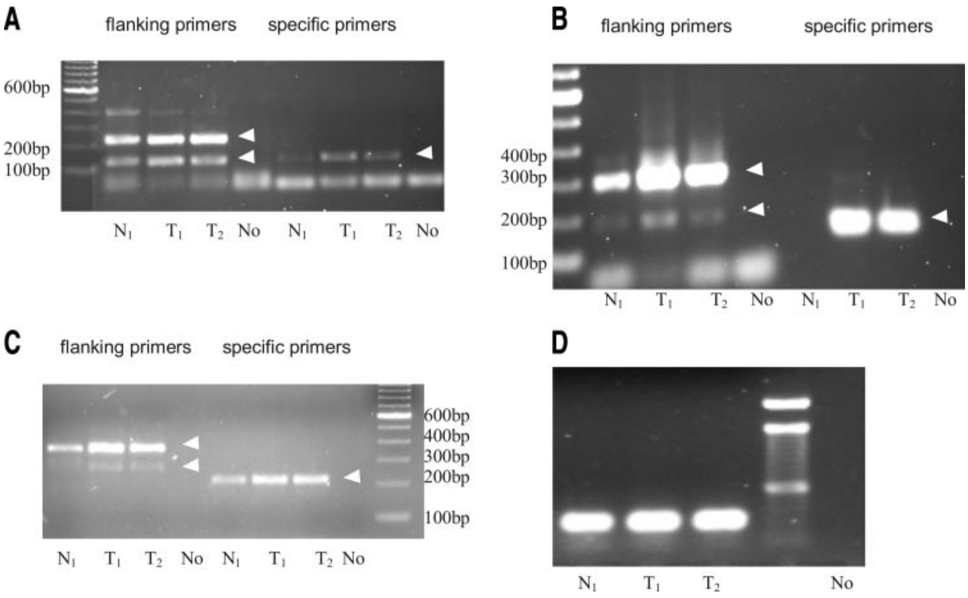

Fig. 2. Experimental validation of 3 brain tumor-associated candidates after the statistical filter. Candidates that passed the statistical filter were randomly chosen to be screened for expression of their tumor-associated exons by RT-PCR on cDNA from normal tissues and tumor cell lines. Three primers 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. N_1, normal whole brain tissue; T_1, T98G glioblastoma cell line; T_2, 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*.

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.

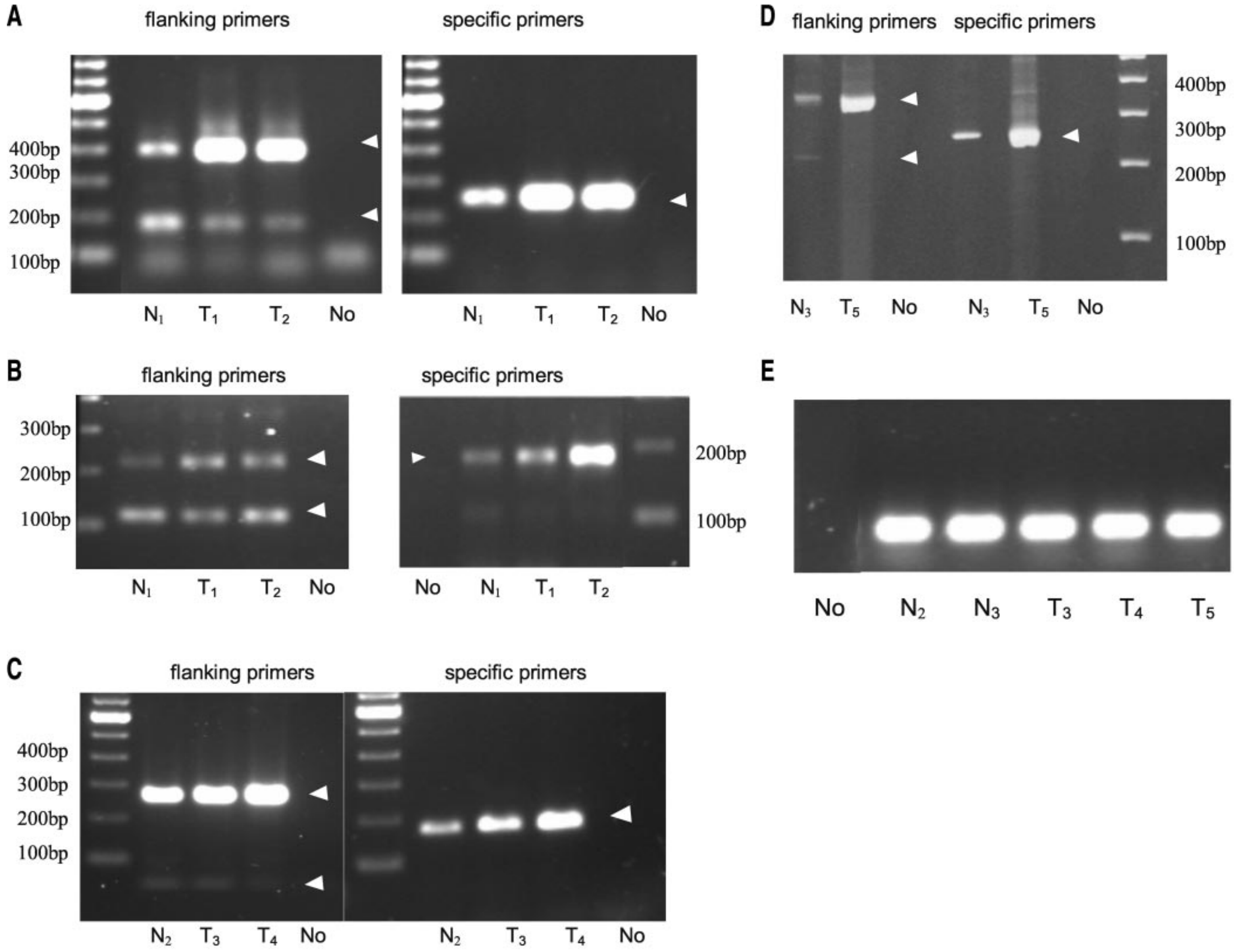

Fig. 3. Experimental validation after the serial analysis of gene expression (SAGE) filter. As in Fig. 2, candidates that passed both the statistical filter and the SAGE filter were randomly chosen to be screened for expression of their tumor-associated exons by RT-PCR on cDNA from normal tissues and tumor cell lines. Three primers were designed for each candidate: 2 on the exons flanking the candidate tumor-associated exon and 1 on the exon itself. Candidate exons corresponded to the following genes: "*Delta Tubulin*" (BC000258; *A*), *Zinc finger protein 585A* (AK074345)-T1 (*B*), *RNA terminal phosphate cyclase-like 1* (BC001025; *C*), and *karyopherin* (*importin*) *beta 1* (NM_002265; *D*). N1, normal whole brain tissue; T_1, T98G glioblastoma cell line; T_2, A172 glioblastoma cell line; N_2, normal lung tissue; T_3, H1155 lung tumor cell line; T_4, H358 lung tumor cell line; N_3, normal colon tissue; T_5, SW480 colon tumor cell line; No, no DNA control. The products from the second reaction (using the primers annealing to the exon itself) showed overexpression of the candidate variant in the respective tumor cell line. However, for the primers flanking the candidate exon, it is shown that only the variant expressing the candidate exon was overexpressed in the tumor cell lines. *A*–*C*: agarose gels. *D*: a silver-stained polyacrylamide gel (the variant skipping the candidate exon for this gene was only visualized this way due to its very low expression level). Amplification of *GAPDH* as a positive control is shown for all samples in *E*.

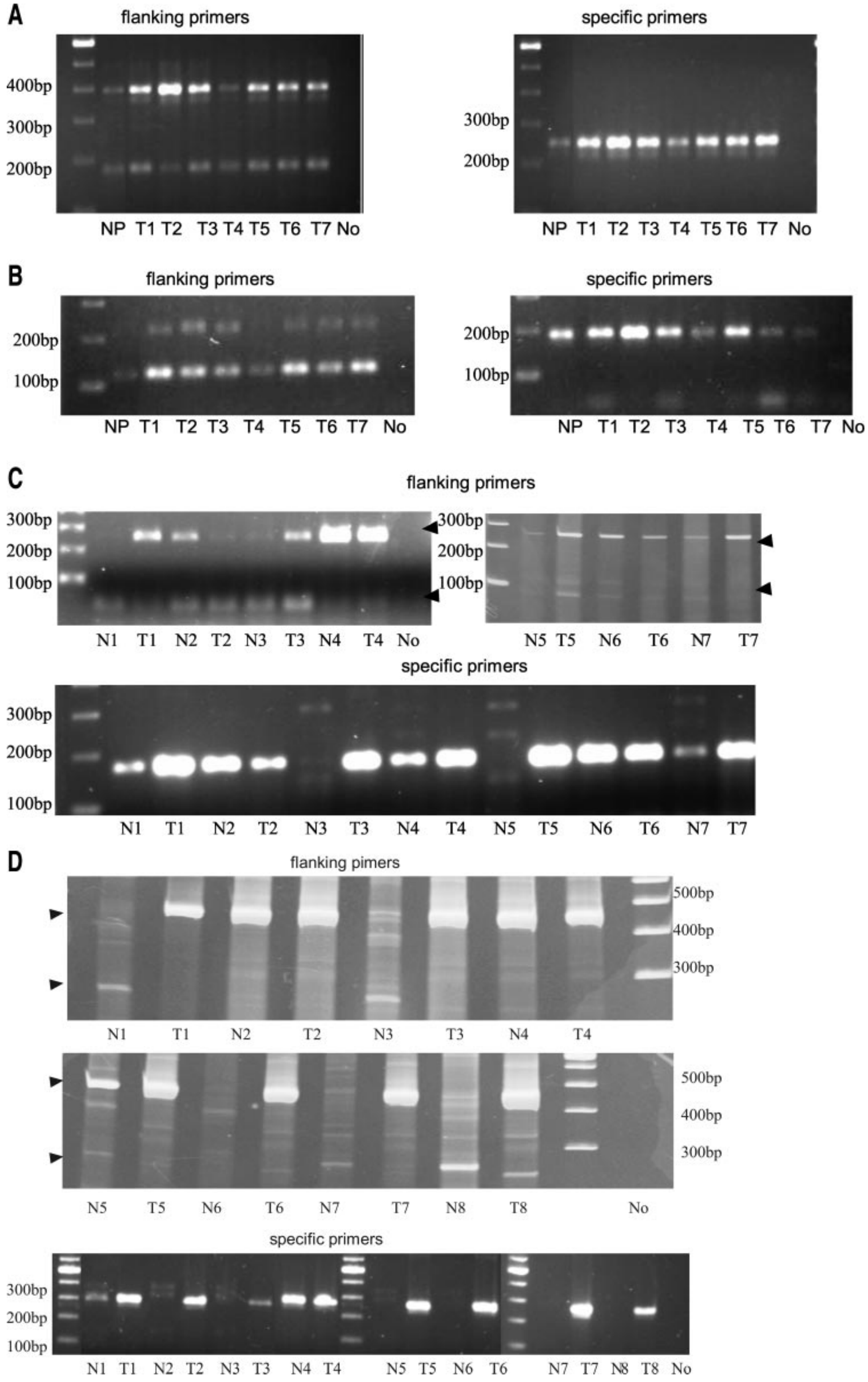

Fig. 4. Experimental validation in patient tumor samples. The 4 candidate genes that were positively validated in tumor cell lines were further validated by RT-PCR on cDNA from patient tumor samples using the same primers and conditions as before. For brain tissue, we used 7 tissue samples from glioblastoma patients and compared those with commercially obtained normal pool brain RNA (Clontech Laboratories). For lung and colon, we compared RNA from paired normal/tumor patient samples. For 3 of the 4 candidates, we obtained the same expression pattern as in cell lines in at least 2 patient samples. *A*: *Delta Tubulin* (BC000258) was validated in brain samples. The products from the reaction using the primers annealing to the exon itself showed overexpression of the candidate variant in 6 of the 7 patient samples. The primers flanking the candidate exon show that only the variant expressing the candidate exon was overexpressed in these patients. NB, pool of normal whole brain tissue; T_1–T_7, 7 different brain tumor patient samples; No, no DNA control. *B*: *Zinc finger protein 585A* (AK074345) was validated in brain samples. The products from the reaction using the primers annealing to the exon itself showed overexpression of the candidate variant in 4 of the 7 patient samples. However, the primers flanking the candidate exon showed that both the variant expressing the candidate exon and the prototype were overexpressed in these patients. NB, pool of normal whole brain tissue; T_1–T_7, 7 different brain tumor patient samples; No, no DNA control. *C*: *RNA terminal phosphate cyclase-like 1* (BC001025) was validated in paired lung samples. The products from the reaction using the primers annealing to the exon itself showed overexpression of the candidate variant in 5 of the 7 patient samples. The primers flanking the candidate exon showed that only the variant expressing the candidate exon was overexpressed in at least 2 of these patients (*patients 1* and *7*). N_1–N_7 and T_1–T_7, different paired normal/tumor lung samples, respectively; No, no DNA control. *D*: *karyopherin* (*importin*) *beta 1* (NM_002265) was validated in paired colon samples. The products from the reaction using the primers annealing to the exon itself showed overexpression of the candidate variant in 7 of 8 patient samples. The primers flanking the candidate exon showed that only the variant expressing the candidate exon was overexpressed in at least 5 of these patients (*patients 1*, *3*, *5*, *7*, and *8*). N_1–N_8 and T_1–T_8, different paired normal/tumor colon samples, respectively; No, no DNA control. Amplification of *GAPDH* as a positive control is shown in Supplemental Fig. S2.

Functional classification of the final candidate list. 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

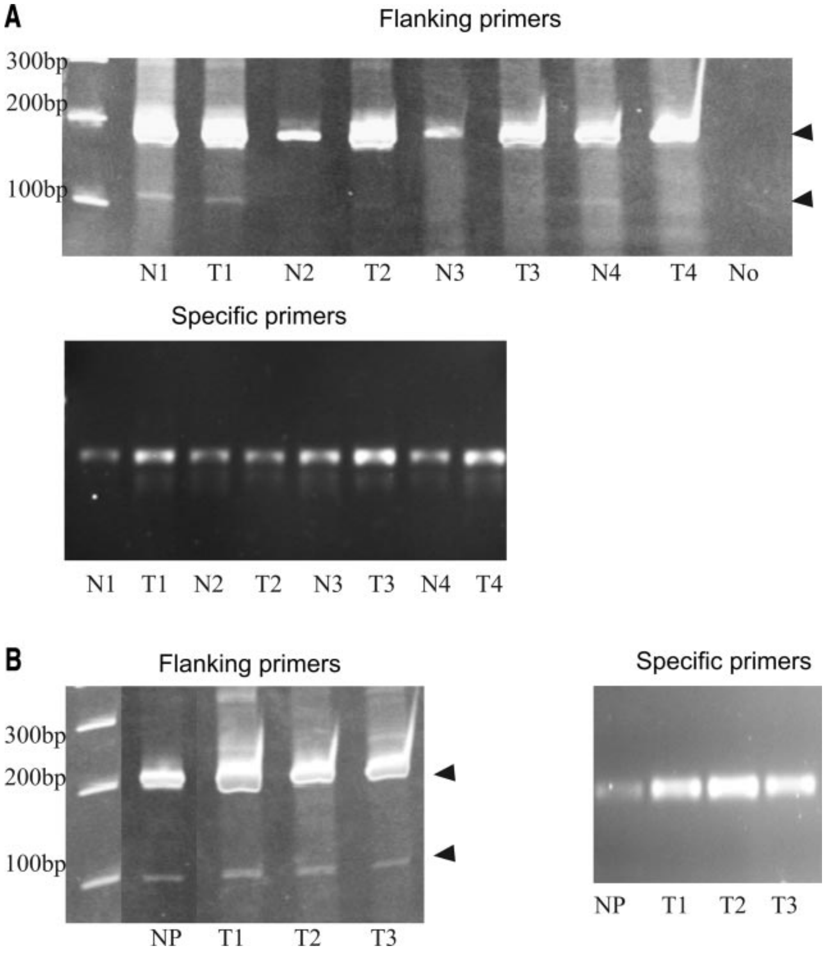

Fig. 5. We evaluated the expression pattern of 2 more candidates that were originally negative in the validation using tumor cell lines. *A*: the gene *NADH dehydrogenase* (*ubiquinone*) *Fe-S protein 2, 49-kDa* (*NADH-coenzyme Q reductase*) (BC001456), was validated in paired lung samples. N_1–N_4 and T_1-T_4, same paired lung normal/tumor patient samples, respectively, as in Fig. 4*C*; 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; T_1–T_3, different prostate tumor patient samples; No, no DNA control.

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)*

Gene	Overlap of Candidates Before SAGE Filter	Overlap of Candidates After SAGE Filter
NME1	+	−
CDC25C	+	−
DVL1	+	−
ERCC1	+	−
GSS	+	−
GTF3C1	+	+
IRAK1	+	−
NKTR	+	−
POLB	+	+
RAD51	+	−
SHC1	+	−
ST5	+	+
TNFRSF1A	+	−

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 *RAB1A* (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.

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

Gene	Overlap of Candidates Before SAGE Filter	Overlap of Candidates After SAGE Filter
BZW2	+	+
CCT3	+	−
CGI-31	+	−
F11R	+	−
HGD	+	+
HLA-DMB	+	+
HNRPF	+	−
HNRPK	+	−
KIF2C	+	+
MBP	+	+
MED8	+	+
MGC11257	+	+
MORF4L2	+	−
NGLY1	+	+
NUDE1	+	+
SAD1	+	−
SMUG1	+	−
SPC18	+	−
TPM1	+	−
WDR4	+	+
WRB	+	−

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 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-Slager 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

1. **Baudry D, Hamelin M, Cabanis MO, Fournet JC, Tournade MF, Sarnacki S, Junien C, and Jeanpierre C.** WT1 splicing alterations in Wilms' tumors. *Clin Cancer Res* 6: 3957–3965, 2000.
2. **Beissbarth T and Speed TP.** GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* 20: 1464–1465, 2004.
3. **Black DL.** Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* 72: 291–336, 2003.
4. **Boguski MS, Lowe TMJ, and Tolstoshev CM.** dbEST—database for "expressed sequence tags." *Nat Genet* 4: 332–333, 1993.
5. **Boon K, Osorio EC, Greenhut SF, Schaefer CF, Shoemaker J, Polyak K, Morin PJ, Buetow KH, Strausberg RL, de Souza SJ, and Riggins GJ.** An anatomy of normal and malignant gene expression. *Proc Natl Acad Sci USA* 99: 11287–11292, 2002.
7. **Brett D, Hanke J, Lehmann G, Haase S, Delbruck S, Krueger S, Reich J, and Bork P.** EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett* 474: 83–86, 2000.
8. **Caballero OL, de Souza SJ, Brentani RR, and Simpson AJG.** Alternative spliced transcripts as cancer markers. *Dis Markers* 17: 67–75, 2001.
9. **Chirgwin JM, Przybyla AE, MacDonald RJ, and Rutter WJ.** Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. *Biochemistry* 18: 5294–5299, 1979.
10. **Cragg MS, Chan HTC, Fox MD, Tutt A, Smith A, Oscier DG, Hamblin TJ, and Glennie MJ.** The alternative transcript of CD79b is overexpressed in B-CLL and inhibits signaling for apoptosis. *Blood* 100: 3068–3076, 2002.
11. **Croft L, Schandorff S, Clark F, Burrage K, Arctander P, and Mattick JS.** ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nat Genet* 24: 340–341, 2000.
12. **Galante PA, Sakabe NJ, Kirschbaum-Slager N, and de Souza SJ.** Detection and evaluation of intron retention events in the human transcriptome. *RNA* 10: 757–765, 2004.
13. **Ge K, DuHadaway J, Du W, Herlyn M, Rodeck U, and Prendergast GC.** Mechanism for elimination of a tumor suppressor: aberrant splicing of a brain-specific exon causes loss of function of Bin1 in melanoma. *Proc Natl Acad Sci USA* 96: 9689–9694, 1999.
14. **Hide WA, Babenko VN, van Heusden PA, Seoighe C, and Kelso JF.** The contribution of exon-skipping events on chromosome 22 to protein coding diversity. *Genome Res* 11: 1848–1853, 2001.
15. **Hui L, Zhang X, Wu X, Lin Z, Wang Q, Li Y, and Hu G.** Identification of alternatively spliced mRNA variants related to cancers by genome-wide ESTs alignment. *Oncogene* 23: 3013–3023, 2004.

15a.**Human Cancer Genome Project/Cancer Genome Anatomy Project Annotation Consortium; Human Cancer Genome Project Sequencing Consortium.** 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, 2003.

16. **International Human Genome Sequencing Consortium.** Initial sequencing and analysis of the human genome. *Nature* 409: 860–921, 2001.
17. **Kan Z, States D, and Gish W.** Selecting for functional alternative splices in ESTs. *Genome Res* 12: 1837–1845, 2002.
18. **Kelso J, Visagie J, Theiler G, Christoffels A, Bardien S, Smedley D, Otgaar D, Greyling G, Jongeneel CV, McCarthy MI, Hide T, and Hide W.** eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res* 13: 1222–1230, 2003.
19. **Krawczak M, Reiss J, and Cooper DN.** The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum Genet* 90: 41–54, 1992.
20. **McKeown M.** Alternative mRNA splicing. *Annu Rev Cell Biol* 8: 133–155, 1992.
21. **Michalik L, Desvergne B, and Wahli W.** Peroxisome-proliferator-activated receptors and cancers: complex stories. *Nat Rev Cancer* 4: 61–70, 2004.
22. **Mironov AA, Fickett JW, and Gelfand MS.** Frequent alternative splicing of human genes. *Genome Res* 9: 1288–1293, 1999.
23. **Modrek B, Resch A, Grasso C, and Lee C.** Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res* 29: 2850–2859, 2001.
24. **Modrek B and Lee C.** A genomic view of alternative splicing. *Nat Genet* 30: 13–19, 2002.
25. **Nagoshi RN, McKeown M, Burtis KC, Belote JM, and Baker BS.** The control of alternative splicing at genes regulating sexual differentiation in *D. melanogaster*. *Cell* 53: 229–236, 1988.
26. **Naor D, Sionov RV, and Ish-Shalom D.** CD44: structure, function, and association with the malignant process. *Adv Cancer Res* 71: 241–319, 1997.
27. **Osorio EC, de Souza JE, Zaiats AC, de Oliveira PS, and de Souza SJ.** pp-Blast: a "pseudo-parallel" Blast. *Braz J Med Biol Res* 36: 463–464, 2003.
28. **Sakabe NJ, de Souza JES, Galante PAF, de Oliveira PSL, Passetti F, Brentani H, Osorio EC, Zaiats AC, Leerkes MR, Kitajima JP, Brentani RR, Strausberg RL, Simpson AJG, and de Souza SJ.** ORESTES are enriched in rare exon usage variants affecting the encoded proteins. *CR Biol* 326: 979–985, 2003.
29. **Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, Rodriguez-Tome P, Aggarwal A, Bajorek E, Bentolila S, Birren BB, Butler A, Castle AB, Chiannilkulchai N, Chu A, Clee C, Cowles S, Day PJ, Dibling T, Drouot N, Dunham I, Duprat S, East C, Edwards C, Fan JB, Fang N, Fizames C, Garrett C, Green L, Hadley D, Harris M, Harrison P, Brady S, Hicks A, Holloway E, Hui L, Hussain S, Louis-Dit-Sully C, Ma J, MacGilvery A, Mader C, Maratukulam A, Matise TC, McKusick KB, Morissette J, Mungall A, Muselet D, Nusbaum HC, Page DC, Peck A, Perkins S, Piercy M, Qin F, Quackenbush J, Ranby S, Reif T, Rozen S, Sanders C, She X, Silva J, Slonim DK, Soderlund C, Sun WL, Tabar P, Thangarajah T, Vega-Czarny N, Vollrath D, Voyticky S, Wilmer T, Wu X, Adams**

MD, Auffray C, Walter NA, Brandon R, Dehejia A, Goodfellow PN, Houlgatte R, Hudson JR Jr, Ide SE, Iorio KR, Lee WY, Seki N, Nagase T, Ishikawa K, Nomura N, Phillips C, Polymeropoulos MH, Sandusky M, Schmitt K, Berry R, Swanson K, Torres R, Venter JC, Sikela JM, Beckmann JS, Weissenbach J, Myers RM, Cox DR, James MR, Bentley D, Deloukas P, Lander ES, and Hudson TJ. A gene map of the human genome. *Science* 274: 540–546, 1996.

30. **Vandenbroucke II, Vandesompele J, Paepe AD, and Messiaen L.** Quantification of splice variants using real-time PCR. *Nucleic Acids Res* 29: E68, 2001.
31. **Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, and Zhu X.** The sequence of the human genome. *Science* 291: 1304–1351, 2001.
32. **Wang L, Duke L, Zhang PS, Arlinghaus RB, Symmans WF, Sahin A, Mendez R, and Dai JL.** Alternative splicing disrupts a nuclear localization signal in spleen tyrosine kinase that is required for invasion suppression in breast cancer. *Cancer Res* 63: 4724–4730, 2003.
33. **Wang Z, Lo HS, Yang H, Gere S, Hu Y, Buetow KH, and Lee MP.** Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. *Cancer Res* 63: 655–657, 2003.
34. **Xie H, Zhu WY, Wasserman A, Grebinskiy V, Olson A, and Mintz L.** Computational analysis of alternative splicing using EST tissue information. *Genomics* 80: 326–330, 2002.
35. **Xu Q and Lee C.** Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. *Nucleic Acids Res* 31: 5635–5643, 2003.
36. **Zhang B, Schmoyer D, Kirov S, and Snoddy J.** GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics* 5: 16, 2004.
37. **Zhang Z, Schwartz S, Wagner L, and Miller W.** A greedy algorithm for aligning DNA sequences. *J Comput Biol* 7: 203–214, 2000.
38. **Zhu J, Shendure J, Mitra RD, and Church GM.** Single molecule profiling of alternative pre-mRNA splicing. *Science* 301: 836–838, 2003.