

Identification of Differentially Regulated Splice Variants and Novel Exons in Glial Brain Tumors Using Exon Expression Arrays

Pim J. French,¹ Justine Peeters,² Sebastiaan Horsman,² Elza Duijm,¹ Ivar Siccama,¹ Martin J. van den Bent,¹ Theo M. Luider,¹ Johan M. Kros,³ Peter van der Spek,² and Peter A. Sillevs Smitt¹

Departments of ¹Neurology, ²Bioinformatics, and ³Pathology, Erasmus MC, Rotterdam, the Netherlands

Abstract

Aberrant splice variants are involved in the initiation and/or progression of glial brain tumors. We therefore set out to identify splice variants that are differentially expressed between histologic subgroups of gliomas. Splice variants were identified using a novel platform that profiles the expression of virtually all known and predicted exons present in the human genome. Exon-level expression profiling was done on 26 glioblastomas, 22 oligodendrogliomas, and 6 control brain samples. Our results show that Human Exon arrays can identify subgroups of gliomas based on their histologic appearance and genetic aberrations. We next used our expression data to identify differentially expressed splice variants. In two independent approaches, we identified 49 and up to 459 exons that are differentially spliced between glioblastomas and oligodendrogliomas, a subset of which (47% and 33%) were confirmed by reverse transcription-PCR (RT-PCR). In addition, exon level expression profiling also identified >700 novel exons. Expression of ~67% of these candidate novel exons was confirmed by RT-PCR. Our results indicate that exon level expression profiling can be used to molecularly classify brain tumor subgroups, can identify differentially regulated splice variants, and can identify novel exons. The splice variants identified by exon level expression profiling may help to detect the genetic changes that cause or maintain gliomas and may serve as novel treatment targets. [Cancer Res 2007;67(12):5635–42]

Introduction

Diffuse gliomas are the most common primary central nervous system tumors in adults (1, 2), and it is estimated that 43,800 new patients are diagnosed in 2005 with a primary brain tumor in the United States.⁴ Based on their histologic appearance, gliomas can be divided into astrocytic tumors, pure oligodendroglial tumors, and mixed oligoastrocytic tumors according to standard WHO classification (3). Despite advances in neurosurgery, chemotherapy, and radiotherapy, the prognosis for most glioma patients remains dismal (4, 5).

There is strong evidence that aberrant splice isoforms are involved in the initiation and/or progression of glial brain tumors (6). For example, glioblastomas with epidermal growth factor

receptor (EGFR) amplification frequently (32 of 48) express EGFRvIII, a tumor-specific, ligand-independent, constitutively active isoform of the EGFR that lacks exons 2 to 7 (7). Expression of this splice variant can induce glioma formation in mice (8) and is associated with response to EGFR kinase inhibitors in human (9). Other, (activating) aberrant *EGFR* splice variants are also frequently observed in gliomas (10). In addition, many nervous system cancer-related splice variants were identified using a gene-centric (11–16) or bioinformatical approach screening public domain databases (17).

Because aberrant splice isoforms are involved in the initiation and/or progression of glial brain tumors, we initiated a screen to identify splice variants expressed in gliomas. Our screen was done by profiling the expression of virtually all known and predicted exons in the human genome (1.4 million). Splice variants were then calculated from the expression level of exons relative to its transcript. Our results indicate that exon level expression profiling can classify brain tumor subgroups based on their histologic appearance, can identify differentially regulated splice variants, and can identify novel exons.

Materials and Methods

Samples. All glioma samples were derived from patients treated within the Erasmus MC. Patient data, histologic diagnosis, and chromosomal aberrations are summarized in Supplementary Table S1. Samples were collected immediately after surgical resection, snapped frozen, and stored at -80°C . All samples were visually inspected on 5- μm H&E-stained frozen sections by the neuropathologist (J.M.K.). We selected 48 glioma samples including (a) classic oligodendrogliomas with loss of heterozygosity (LOH) on 1p and 19q ($n = 22$, of which 20 WHO grade III and 2 WHO grade II; ref. 3); (b) primary glioblastoma with EGFR amplification ($n = 18$); and (c) secondary glioblastoma without EGFR amplification ($n = 8$). Six control brain samples from patients with no history of neurologic disease were also included. All but one sample (GBM 77) contained >70% tumors. Tissue adjacent to the inspected sections was subsequently used for nucleic acid isolation. Microsatellite analysis on 1p and 19q and amplification of the EGFR were done as described (18).

Nucleic acid isolation, cDNA synthesis, and array hybridization. Total RNA and genomic DNA was isolated from 20 to 40 cryostat sections of 40- μm thickness (50–100 mg) using Trizol (Invitrogen) according to the manufacturer's instructions (see also ref. 18). Total RNA was then further purified on RNeasy mini columns (Qiagen). RNA quality was assessed on a Bioanalyser (Agilent). High-quality RNA (i.e., RNA integrity number >7.0; ref. 19) was used for our experiments. rRNA reduction, first round double-strand-cDNA synthesis, cRNA synthesis, second round single-strand (ss)-cDNA synthesis, ss-cDNA fragmentation, and labeling was done according to the Affymetrix GeneChip Whole-Transcript Sense Target-Labeling Assay manual. Affymetrix Human Exon 1.0 ST microarrays were hybridized overnight with 5- μg biotin-labeled ss-cDNA.

Note: Supplementary data for this article are available at Cancer Research Online (<http://cancerres.aacrjournals.org/>).

Requests for reprints: Pim J. French, Department of Neurology, Erasmus MC, P.O. Box 2040, 3000CA Rotterdam, the Netherlands. Phone: 311-040-88333; Fax: 11-31-10-4088365; E-mail: p.french@erasmusmc.nl.

©2007 American Association for Cancer Research.
doi:10.1158/0008-5472.CAN-06-2869

⁴ CBTRUS 2004–2005 statistical report (www.cbtrus.org).

Data analysis. Signal intensity estimate and *P* value for each probe set were extracted from the arrays in Affymetrix ExACT 1.0 software using the PLIER and DABG algorithm, respectively. PLIER expression data were normalized using the quantile method in *R* statistical software v2.2.1. DABG *P* values allow calculation of false positive and negative probe sets at various PLIER expression level cutoff values. The results are summarized in Supplementary Fig. S1 and show that a PLIER expression level of 30 is close to the cutoff that results in the least amount of falsely called probe sets at DABG *P* values of <0.05. A higher cutoff level close to PLIER expression 70 seems to result in the least amount of falsely called probe sets at the more stringent DABG *P* value of <0.01. All values were then imported into Omniviz v3.9 (Omniviz) software for further analysis. For each probe set, the geometric mean of the hybridization intensities of all samples from the

patients was calculated with expression values of <30 set to 30 (close to the optimal cutoff with least amount of falsely called probe sets at DABG *P* value of <0.05).

The expression level of each probe set in every sample was determined relative to the geometric mean and logarithmically transformed (base 2 of scale) to ascribe equal weight to gene expression levels. Deviation from the geometric mean reflects differential probe set expression. Pearson's correlation plots were generated using all probe sets that differed 4-fold from the geometric mean in at least one sample (97175 probe sets in total, Fig. 1) or with DABG *P* < 0.01 in at least five samples (yielding virtually identical similar results, data not shown). Ordering of samples is done according to the algorithm present in Omniviz software as described (20). This method reveals patterns of homologous samples based on Pearson's

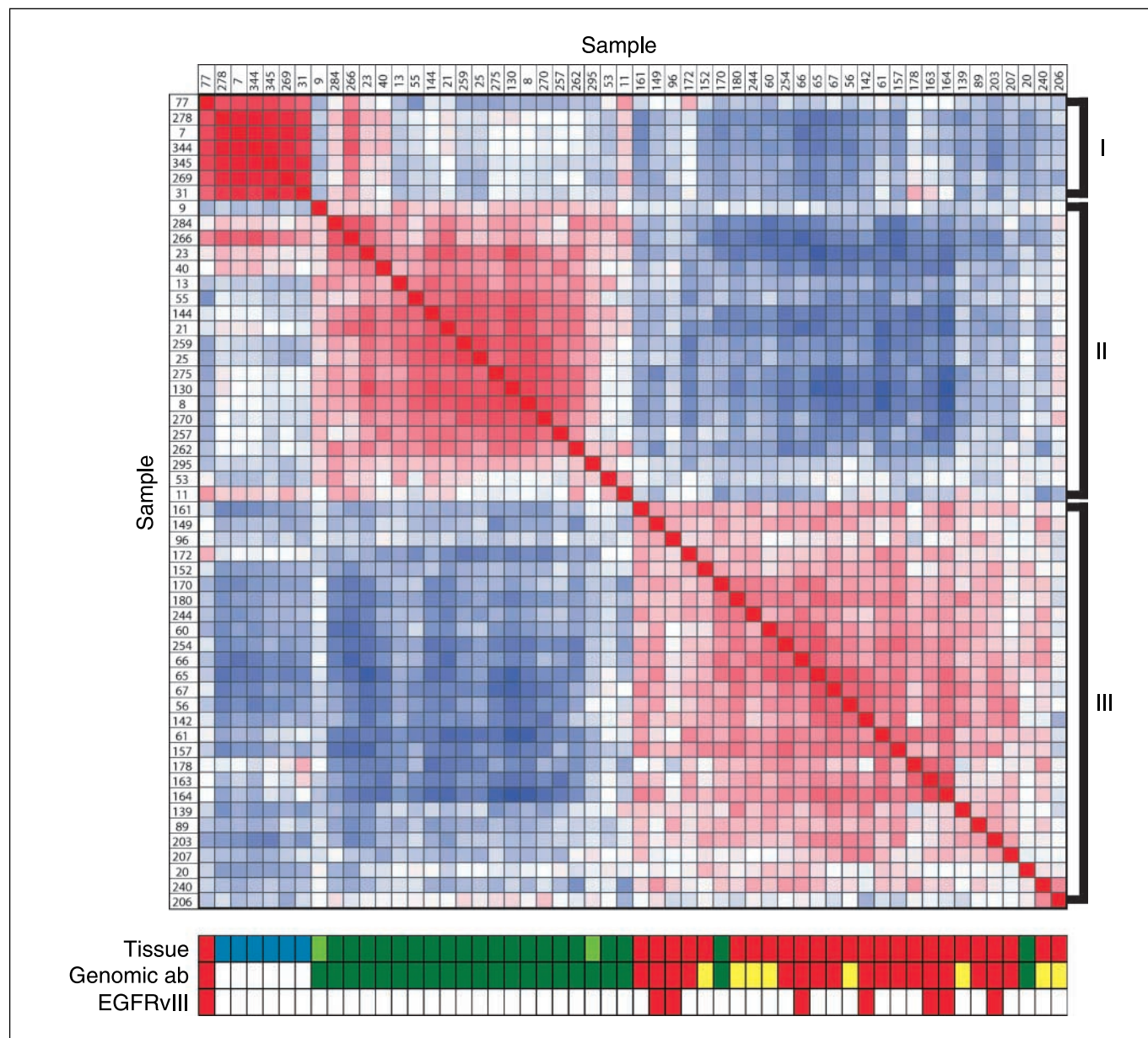


Figure 1. Correlation plot of all samples. Samples are plotted against each other as Pearson's correlation to determine the degree of similarity based on expressed exons. All exons with 4-fold expression difference from the geometric mean are included in the clustering. *Red*, high correlation; *blue*, low correlation. Below the correlation plot is a graphic representation of histologic and patient data. *Tissue.* Origin of sample: ■ control cortex; ■ anaplastic oligodendroglioma (WHO grade III); ■ oligodendroglioma (WHO grade II); and ■ glioblastoma. *Genomic aberrations.* Genomic aberrations of the sample: □ control sample; ■ LOH on 1p and 19q, no amplification of EGFR; ■ no LOH on 1p and 19q but amplification of EGFR; ■ no LOH on 1p and 19q, no amplification of EGFR. *EGFRVIII:* expression of EGFRVIII as determined by RT-PCR: □ no expression; ■ expression. Subgroups identified by Pearson's correlation plot (*right*; I-III).

correlation. The ordering algorithm sorts all samples into correlated blocks through an iterative process and starts with the most highly correlated pair of samples. Each sample is joined to a block, resulting in a correlation trend within a block. The most correlated samples are at the center of each block. The blocks are then positioned along the diagonal of the plot in a similar ordered manner.

Splice variant detection. We used pattern-based correlation (PAC) as an algorithm to identify differentially regulated splice variants. PAC predicts the expression of a probe set in a given sample by the product of its metaprobe set level (a metaprobe set is a collection of probe sets that belong to the same transcript; the metaprobe set level is the calculated transcript level based on the expression level of these probe sets) and the probe set/transcript ratio of all samples:

$$\text{Exp}_{a,c} = \text{Tr}_{b,c} \cdot \text{Ex}_{\text{ave}-a} / \text{Tr}_{\text{ave}-b}$$

where $\text{Exp}_{a,c}$ is the predicted expression of probe set a in sample c, $\text{Tr}_{b,c}$ is the calculated metaprobe set level of transcript b (of which probe set a is part) in sample c, $\text{Ex}_{\text{ave}-a}$ is the measured expression average of probe set a in all samples, and $\text{Tr}_{\text{ave}-b}$ is the expression average of transcript b in all samples. In absence of alternative splicing or when a similar ratio of alternative splicing is observed in all samples, the predicted expression value should be identical to the measured PLIER expression levels:

$$\text{Exm}_{a,c} - \text{Exp}_{a,c} = 0$$

where $\text{Exm}_{a,c}$ is the measured PLIER expression data from the array. Any deviation from 0 in this formula is a predictor for alternative splicing: negative values predict the exon is spliced out in a given sample; positive values predict the exon is spliced in. PAC values were calculated using \log^2 transformed expression data.

Because splice variant detection requires an accurate estimation of metaprobe sets, we used two independent approaches to calculate metaprobe set levels. The first metaprobe set levels were calculated using ExACT 1.0 software based on probe sets determined by Affymetrix. The second metaprobe set calculations required two iterations: We first determined metaprobe set levels by averaging all probe sets with PLIER expression levels >30 , >50 , or >80 . We next hypothesized that differentially spliced exons will result in a metaprobe set level that is lower than when calculated using constitutive exons only. For example, an exon that is spliced out in subgroup A can reduce its metaprobe set level so that constitutive exons are identified as exons that are differentially spliced-out in subgroup B. Therefore, transcript levels should be calculated only using constitutively incorporated (i.e., not differentially spliced between defined subgroups) exons. We defined those constitutive exons (probe sets) as those that are highly correlated (correlation coefficient >0.7 , >0.8 , or >0.9) with the first round transcript calculations. A total of five metaprobe set calculations were done using cutoff values: (a) PLIER 50, correlation 0.8; (b) PLIER 30, correlation 0.8; (c) PLIER 80, correlation 0.8; (d) PLIER 50, correlation 0.7; and (e) PLIER 50, correlation 0.9. This two-step metaprobe set calculation not only excludes differentially spliced exons but also excludes "nonlinear" probe sets (probe sets that are outside the linear detection range of arrays) and "a-specific" probe sets (probe sets that bear no relation to its transcript).

Statistical analysis was done using standard *t* tests. Identical filtering and statistical analysis was done on 10 randomized groups to test for type I errors and estimate the false-discovery rate.

Reverse transcription-PCR. Candidate differentially regulated splice variants identified by PAC analysis were analyzed by reverse transcription-PCR (RT-PCR) to confirm differential regulation. All RT-PCR experiments were done on cDNA that was independently reverse transcribed from the cDNA that was used for array hybridization. rRNA (0.5 μg)-depleted (ribominus RNA) total RNA (the remainder of RNA that was used for array hybridization) was reverse transcribed for 1 h at 42°C in the presence of 200 units of Superscript II, 50 ng T7-(N)₆ primers, 0.5 mmol/L deoxy-nucleotide triphosphates, 10 mmol/L DTT, and RNase inhibitor. Primers

were designed using Primer3⁵ and are listed in Supplementary Table S2. Amplified PCR products from novel exon analysis were sequence verified using the Big Dye Terminator Cycle Sequencing kit (Applied Biosystems). Reactions were run on an ABI 3100 genetic analyzer.

Results

Human Exon arrays performance and unsupervised clustering. In this study, we performed exon level expression profiling to identify differentially expressed splice variants in glial brain tumors. Profiling was done using Human Exon 1.0 Arrays (Affymetrix), a novel platform that determines the expression of virtually all exons present in the human genome. These arrays are designed to target all well-annotated (RefSeq) exons (*core* exons), less well-characterized exons [e.g., derived from unique EST sequences that are not included in the RefSeq database (*extended* exons)] and all predicted exons (*full* exons) for which no expression data is present in public domain databases. In total, ~1.4 million probe sets (a set of up to four oligonucleotide probes that examines the expression of a single exon) are spotted on Human Exon 1.0 arrays: 284,000 *core*, 523,000 *extended*, and 580,000 *full* probe sets. Multiple probe sets may be directed against the same exon, thus, allowing identification of alternative splice-acceptor or splice-donor sites. Exon arrays also allow calculation of whole-transcript levels based on the expression level of probe sets that belong to the same transcript. Calculated transcript levels are called metaprobe set levels. In our experiments, the DABG significant expression ($P < 0.01$) of $23.7 \pm 4.5\%$ of all 1.4 million probe sets were detected. *Core* exons are detected at higher signal intensities than *extended* and *full* exons (Supplementary Fig. S2). Individual sample performance for all array quality control variables is stated in Supplementary Table S3.

This platform has thus far not been characterized, and we therefore first validated the performance of these arrays using unsupervised clustering analysis. Unsupervised clustering was done using probe sets with PLIER expression levels of >30 that differed 4-fold from the geometric mean in at least one sample (Fig. 1). A first subgroup (I) consists of all control samples and GBM 77, a sample that contained a low amount ($<10\%$) of tumor. A second subgroup (II) consists of most (20 of 22) of the oligodendrogliomas with LOH on 1p and 19q. The final subgroup (III) predominantly (25 of 27) consists of glioblastomas but also includes two oligodendrogliomas with 1p and 19q LOH (OD20 and OD170). Interestingly, OD20 also did not cluster with the majority of oligodendrogliomas with 1p/19q LOH using expression profiling on HU133 plus 2 microarrays (18). Identical subgroups were identified by principle components analysis, using all core probe sets or core metaprobe sets (Supplementary Fig. S2). Unsupervised clustering therefore indicates that exon expression profiling can identify brain tumor subgroups based on their histologic appearance. Our data therefore confirm the observation that histologically defined glioma subgroups are molecularly distinct (for review, see ref. 21) and indicates that, on a global scale, this novel platform performs similar to other expression profiling platforms.

Identification of differentially regulated splice variants. We next examined whether Human Exon arrays can detect glioma subgroup-specific splice variants. The identification of splice variants was done using PAC. PAC values represent a predicted level of expression for each probe set. Therefore, differences

⁵ http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi

between PAC and expression values are indicative for alternative splicing. Negative values predict that the exon is, compared with the other 53 samples, being spliced out. However, PAC requires a complete linearity of all probe sets within a single transcript: if a transcript is up-regulated 2-fold in one subgroup, all of the probe sets that belong to this transcript should be up-regulated exactly 2-fold. Any probe set that does not exhibit this linearity in expression detection (nonlinear probe sets) or bear no correlation whatsoever with its native transcript (a-specific probe sets) will be identified as a false positive differentially spliced candidate. Examples of such nonlinear and a-specific probe sets are shown in Supplementary Fig. S3. Any strategy to identify differentially expressed splice variants therefore requires filtering out nonlinear and a-specific probe sets.

We adopted two independent strategies to identify candidate splice variants that are differentially regulated between oligodendrogliomas and glioblastomas. In the first strategy, we calculated PAC values for every probe set in all samples using metaprobesets predetermined by Affymetrix. For our second strategy, we calculated PAC values using recalculated metaprobeset expression levels (see Materials and Methods) with metaprobeset levels (and subsequent PAC values) derived at varying PLIER expression level and/or correlation coefficient cutoff values. We then aimed to exclude nonlinear and a-specific probe sets using the filtering steps outlined in Fig. 2 and Table 1. These filtering steps resulted in final set of 49 (first strategy) and 254 to 459 candidate differentially

regulated splice variants (second strategy). Table 1 summarizes the results at each step in our strategy to identify candidate splice variants. Supplementary Table S4 contains a list of all candidates.

To estimate the false discovery rate, we randomly assigned a group number to each tissue sample and then repeated the filtering and statistical analysis (Table 1). This scrambling procedure was repeated 10 times and failed to identify any candidate splice variant in the first strategy and 1.8 candidate splice variants (range, 0–7) in the second strategy.

Altering the variables used for metaprobeset calculation often resulted in significant overlap between candidates identified: many candidates identified at cutoff values PLIER 50, and correlation coefficient 0.8 are also found when the PLIER expression cutoff is reduced to 30 (88%), increased to 80 (83%), or the correlation cutoff is reduced to 0.7 (93%). In contrast, increasing the correlation cutoff to 0.9 results in a set of candidates that contains only 50% of the probe sets identified by PLIER 50 correlation 0.8 with 46 additional probe sets identified.

We did RT-PCR using exon spanning primers to confirm the differential expression of candidate splice variants. RT-PCR was done on 15 candidates from the first screen and 21 candidates from the second screen (PLIER 50, correlation 0.8). RT-PCR candidates were randomly selected from the total number of candidates but omitted candidates with alternative 5'- or 3'-end exons. We confirmed 7 of 15 (47%) from the first screen and 7 of 21 (33%) from the second analysis (Fig. 2). Three of the confirmed candidates

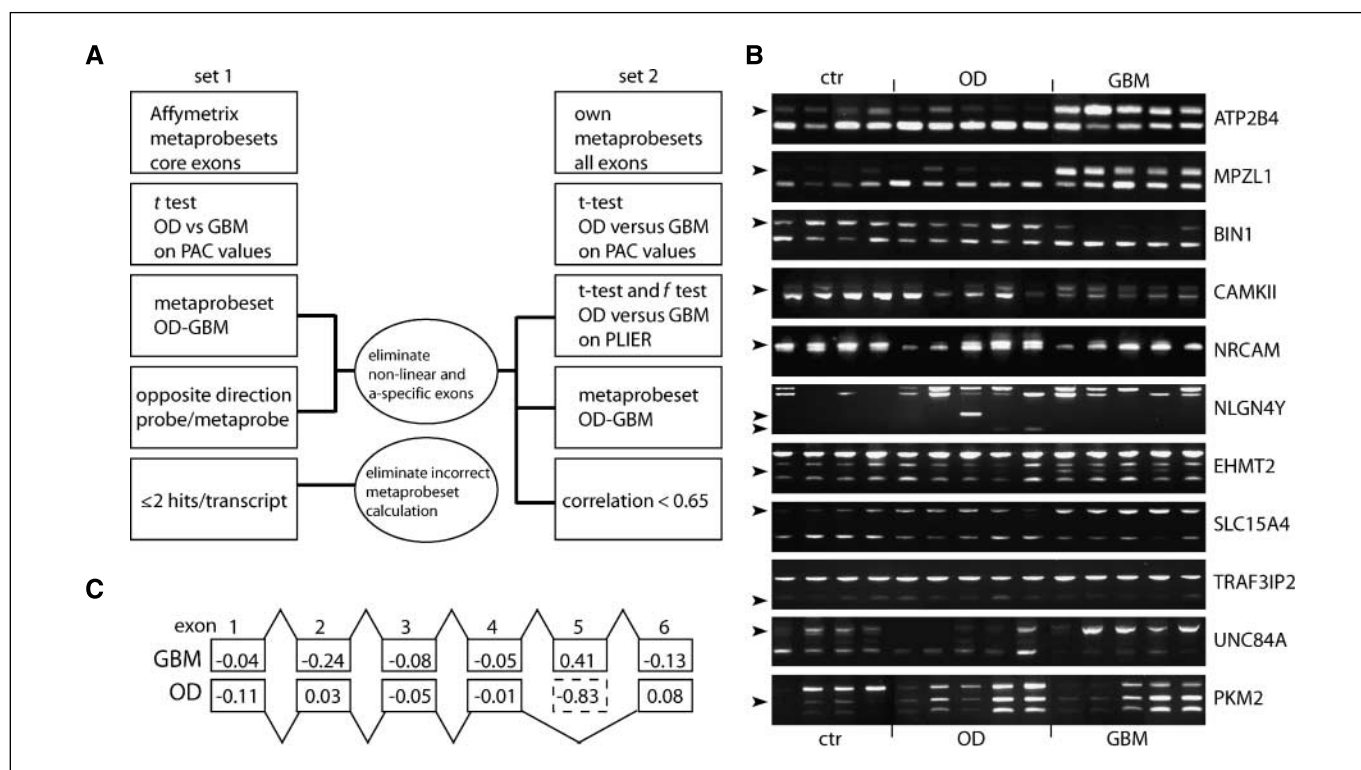


Figure 2. Identification of differentially expressed splice variants. **A**, summary of filtering steps used to identify 49 and 254 to 459 candidate differentially expressed exons, see also Table 1. **B**, RT-PCR of identified candidates using exon-spanning primers. *ATP2B4*, *CaMKII*, *NLGN4Y*, and *UNC84A* were confirmed hits identified in set 1. *BIN1*, *MPZL1*, and *NRCAM* were confirmed hits from sets 1 and 2. Other candidates were confirmed from set 2. In *NLGN4Y*, an exon 5' to the exon identified by PAC also shows alternative splicing, although this exon (exon 3) does not seem to be differentially expressed between oligodendrogliomas and glioblastomas. *Top arrowhead*, transcripts lacking only exon 4; *bottom arrowhead*, transcripts lacking both exons 3 and 4. RT-PCR products of *PKM2* were digested with *pstI*: the differentially spliced exon is mutually exclusive with a 5' exon of identical length. This exon however does not contain a *pstI* restriction site. **C**, model of alternative splicing of *MPLZ1*. In oligodendrogliomas, exon 5 is spliced out, identified by PAC analysis, and confirmed by RT-PCR. PAC values are stated in the represented exons. *OD*, oligodendrogliomas; *GBM*, glioblastomas.

Table 1. Filtering steps used to identify candidate differentially expressed exons

First strategy	Probe sets	PAC values	PAC <i>t</i> test	Transcr GMB ~ OD	Diff exp ex-tr	<3 ex/tr		
Affy metaprobesets core	286,000	188,419	7,776	5,934	65	49		
Random sampling	286,000	188,419	16	16	0	0		
Range			6–35	6–35				
FDR			0.20%	0.30%				
Second strategy	Probe sets	PAC values	PAC <i>t</i> test	PLIER <i>t</i> test	PLIER <i>f</i> test	Correlation	Expr.	Overlap
PLIER 50 corr 0.8	1,400,000	622,971	33,580	7,754	4,696	443	414	
Random sampling	1,400,000	622,971	2,012	20	7.5	1.8		
Range			1150–3203	2–56	0–20	0–7		
FDR			6%	0.25%	0.15%	0.40%		
PLIER 50 corr 0.9	1,400,000	338,970	22,190	5,364	3,199	267	254	208/254
PLIER 50 corr 0.7	1,400,000	850,382	52,820	9,826	5,737	492	459	388/459
PLIER 30 corr 0.8	1,400,000	701,535	34,799	7,016	4,034	418	390	360/390
PLIER 80 corr 0.8	1,400,000	567,410	32,294	8,358	4,989	418	397	346/397

NOTE: Our first strategy made use of core exons only using metaprobesets predetermined by Affymetrix. For our second strategy, we calculated PAC values using recalculated metaprobeset expression levels (as outlined in Materials and Methods) with metaprobeset levels and the subsequent PAC values being recalculated at various probe set inclusion criteria. PAC values represent the number of probe sets in which PAC values could be calculated, omitting all probe sets with absent metaprobeset levels. *Transcript GBM ~ OD*: all probe sets in which metaprobeset levels differed <3-fold between oligodendrogliomas and glioblastomas. *Diff exp ex-tr*: remaining candidates were further selected by probe sets in which the direction of expression is differential between probe sets and metaprobesets. If the average probe set level expression in OD > GBM, then the average metaprobeset expression should be OD < GBM and vice versa. This filter is likely to exclude many true positive candidates but will also rigorously exclude most nonlinear and a-specific candidates. *<3 ex/tr*: all probe sets with three or more candidates within a single transcript were excluded because these are likely to be false positive candidates due to incorrect metaprobeset calculation. *Correlation*: probe sets with high correlation between probe set and metaprobeset expression were excluded (correlation coefficient > 0.65). This filter is based on the hypothesis that regulated splice variants are expected to have an exon/transcript correlation that is less than constitutively incorporated exons. *Overlap*: number of candidates that were also identified using PLIER 50, correlation 0.8.

were identified in both analysis; the total number of differentially expressed splice variants equaled 11. All differentially expressed splice variants belonged to the *core* probe set list. Public domain databases (EMSEMBL, UCSC, HOLLYWOOD) also indicated that most (9 of 11) RT-PCR confirmed candidates are subject to alternative splicing. It is possible that the percentage of regulated splice variants is higher than the RT-PCR-confirmed 47% to 33%: rare splice variants or splice variants that show only minor differential regulation may not have been detected by RT-PCR. Nevertheless, our results show that exon level expression profiling can identify splice variants that are differentially regulated between histologically defined subgroups of gliomas.

Identification of novel exons. We finally examined whether Human Exon arrays can be used to identify novel exons. We screened for novel exons using the *full* probe set list (580,000 probe sets) because all *full* exons lack evidence for expression in public domain databases. *Full* probe sets are composed of exons that can be predicted (e.g., based on the presence of consensus splice acceptor and donor sites) and of sequences that are conserved between human, mouse, and rat. Candidate novel exons met the following criteria (see Fig. 3): (a) show significant expression (PLIER expression levels ≥ 5 0); (b) are part of a *core* metaprobeset as many *full* probe sets are part of poorly characterized and single-exon transcripts; and (c) should have a high (>0.8) correlation coefficient with its metaprobeset (i.e., the probe set is highly expressed in those samples in which the metaprobeset is highly expressed). These criteria resulted in a final set of 715 *full* probe sets as candidate novel exons. More candidates are identified using

less stringent criteria (exon/transcript correlation ≥ 0.7 , identifies 1482 *full* exons). *In silico* analysis of the first 158 *full* probe sets confirmed that 127 of 158 (80%) are indeed novel exons; they are not present in the RefSeq database and no spliced EST has thus far been identified. Of the remaining probe sets, 18 of 158 (11%) were incorrectly annotated and are in fact part of a RefSeq gene, and 13 of 158 (8%) were identified as part of (rare) spliced ESTs.

We next used RT-PCR to verify that candidate novel exons are indeed expressed as part of a known gene. Primers were designed to span >2 kb intronic sequence to exclude false positives due to amplification of genomic DNA or pre-mRNA sequence. RT-PCR confirmed the expression of 6 of 9 (67%) *full* exons, for which no expression data is present in public domain databases (Fig. 3B). These PCRs were done using one of the primers within the novel exon. We used direct sequencing to confirm that the novel exons are indeed expressed as part of a known transcript and not due to amplification of a-specific products (Fig. 3C). In all cases, products that contain the (RefSeq) known flanking exons and the novel exon were identified. Furthermore, direct sequencing enabled us to confirm the presence of consensus splice acceptor/donor sequences surrounding the novel exons.

RT-PCR also confirmed the expression of 3 of 3 (100%) *full* exons that, in public domain databases, were part of rare spliced ESTs. All three exons could be identified in all examined samples. For *KDHRBS2* and *DTNA*, RT-PCR was done using exon-spanning primers; for *PDE1C*, RT-PCR was done with the forward primer in the candidate novel exon because the novel exon may represent a novel 5' exon. Identification of transcripts that have incorporated

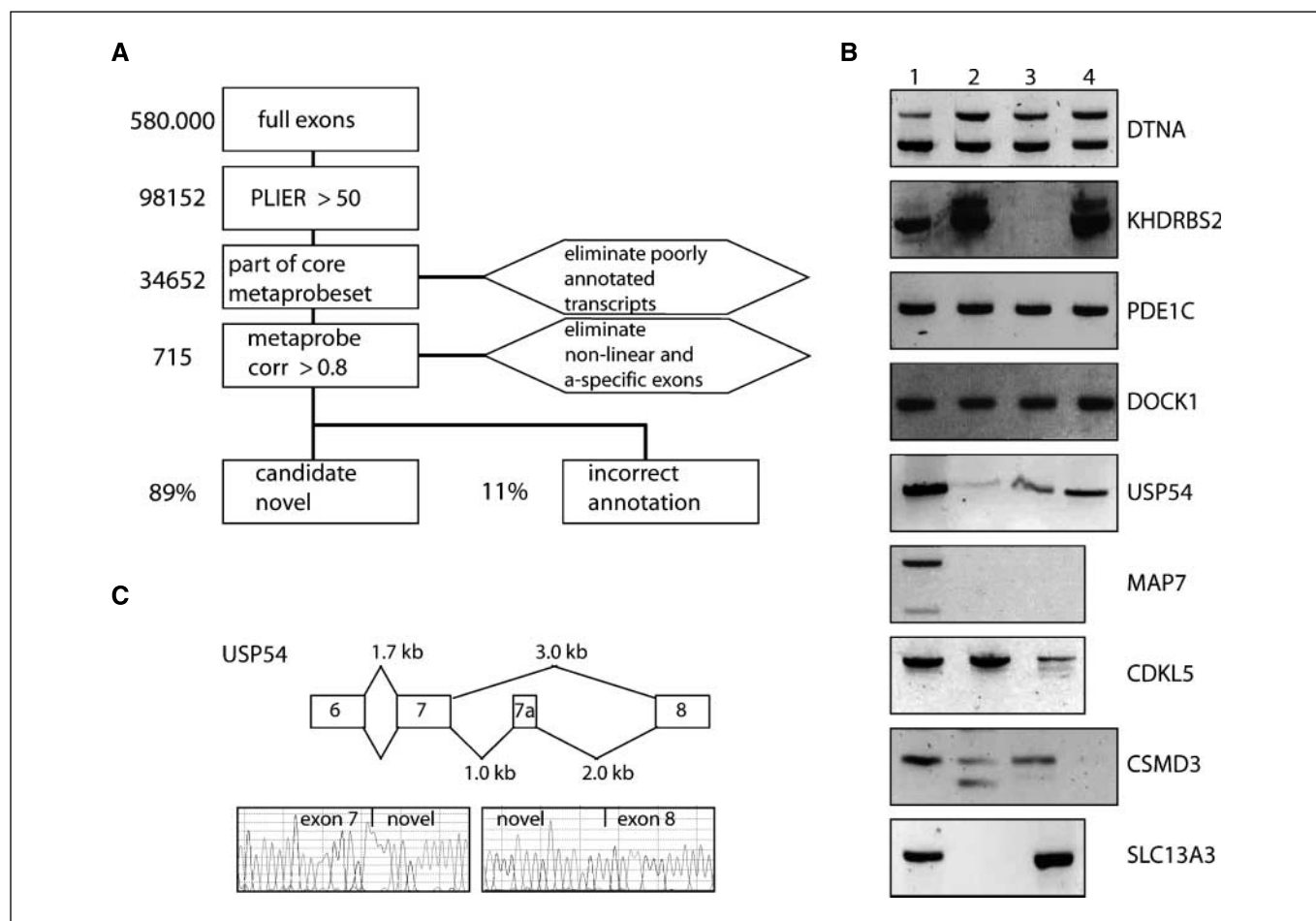


Figure 3. Identification of novel exons by exon level expression profiling. **A**, filtering steps used to identify 715 candidate novel exons. Candidate novel exons are expressed (PLIER) >50 as part of a well-characterized transcript and have a correlation coefficient of >0.8 with its transcript. **B**, RT-PCR of a subset of identified candidates on independent samples (lanes 1–4). *DTNA*, *KHDRBS2*, and *PDE1C* were identified as part of a rare splice variant in public domain databases. Expression of *DTNA* and *KHDRBS2* full exons was confirmed using exon spanning primers, other full exons were confirmed using one primer within the candidate novel exon. Products were sequence verified to exclude a-specific amplifications. **C**, model of splicing of the novel identified exon in *USP54*. Direct sequencing confirmed the presence of the novel exon expressed as part of *USP54*.

the novel exon using exon-spanning primers suggests that a significant percentage of transcripts have incorporated the full exon in adult brain (Fig. 3B).

Discussion

In this study, we did exon level expression profiling on a set of glial brain tumors. To our knowledge, we are among the first to describe the use of Human Exon 1.0 arrays as an expression profiling platform. Our results show that Human Exon arrays can identify subgroups of gliomas based on their histologic appearance and genetic aberrations, can identify differentially expressed splice variants, and can identify novel exons.

The molecular subgroups identified using exon level expression profiling is highly similar to the subgroups that are identified in other studies using 3' biased expression profiling (18, 22–27). Our data therefore confirm the observation that histologically defined glioma subgroups are molecularly distinct (for review, see ref. 21). Furthermore, the similarity in glial tumor classification indicates that, at least on a global scale, this novel platform performs similar to other expression-profiling platforms.

The additional complexity of exon level expression profiling over transcript-level expression profiling is the ability to identify splice variants that are differentially expressed between tumor subgroups. Our data indicate that the identification of differentially expressed splice variants requires rigorous filtering steps to exclude nonlinear and a-specific probe sets. In the two independent approaches adopted by us, we identified 49 and 254 to 459 candidate splice variants that are differentially expressed between OD and GBM. The list of candidates differs significantly between the two approaches. Furthermore, candidates identified by our second approach (recalculated metaprobeset level) are dependent on the inclusion criteria used to recalculate metaprobeset levels. It remains to be determined which variables are optimal for splice variant detection. However, all candidate lists generated by our second approach contain a similar percentage of known splicing events (~12%; range, 10.4–13.8%; see Supplementary Table S4) as determined by screening public domain databases on a subset of candidates.

RT-PCR confirmed the differential regulation of a subset of these candidate splice variants. The select number of differentially expressed splice variants identified by us may reflect the similarity in splice variant expression between OD and GBM. Indeed, a

limited number (591) of differentially expressed splice variants between mouse brain and immune tissue were identified by Ule and coworkers using exon-junction arrays (28). In contrast, experimental evidence exists for the regulated expression of a large number of splice variants: many splice variants show some degree of tissue specificity (29–31). It is therefore also possible that the strong filtering used in this study has led to the identification of only a subset of differentially regulated splice variants.

The differential expression of splice variants between two tumor subtypes may be caused by a differential expression of proteins that regulate alternative splicing. Indeed, a large number of proteins have been identified to play a role in the regulation of alternative splicing (for review, see refs. 32–34). However, the expression of glioma subgroup-specific splice variants may also be a result of genetic changes. For example, glioblastomas with *EGFR* amplifications frequently carry an intragenic deletion of exons 2 through 7, resulting in expression of the tumor specific, constitutively active EGFRvIII isoform (35). Such aberrant splice isoforms have been shown to play a role in the initiation and/or progression of glial brain tumors (6). Identifying glioma-specific splice variants may therefore help identify the causative genetic changes of glial brain tumors.

Apart from exon expression arrays, other techniques have been used to analyze splice variant expression. These include exon-junction arrays (36), RNA-mediated annealing, selection and ligation (37) and digital polony (polymerase colony) exon profiling (38). Recently, arrays containing a combination of exon expression and exon junction probes have also been used to identify alternative splicing events (39, 40). Although all approaches can detect alternative splicing events, many are limited either by screening on a predetermined set of exon-junctions or screening on a per-gene base. Our data shows that exon expression profiling is a suitable alternative for genome-wide screening of regulated splicing events between two distinct subgroups.

Our study has also identified 715 *full* exons that are expressed as part of a well-annotated transcript. *In silico* analysis (screening public domain databases) of a subset of candidates indicated that 80% are indeed novel exons; they are not present in the RefSeq

database and no spliced EST has thus far been identified. We confirmed the expression of ~67%, suggesting a total of ~446 (0.78*0.8*715) novel exons are expressed as part of a well-annotated transcript. Candidates that were not confirmed by RT-PCR (33%) may be falsely identified, for example when the exon array detects unspliced, pre-mRNA species (see e.g., ref. 41). The majority (5 of 6) of RT-PCR confirmed novel exons are expressed in normal adult human brain, indicating they are not aberrant, cancer-specific splice isoforms. Furthermore, most (5 of 6) of the RT-PCR confirmed novel exons result in changes at the protein level: the novel exons are often found within the protein coding region.

Many of the *full* probe sets on the Human Exon arrays are based on evolutionary sequence conservation between human, mouse, and rat. Other studies have also found novel exons based on such sequence conservation. For example, ~150 candidate novel human exons were identified in a screen based on the expression of ESTs in mouse/rat (42). Furthermore, a bioinformatical approach using sequence conservation has identified up to 2,300 novel, rodent-specific exons (43). In a separate study, bioinformatical analysis based on exon expression profiles from adult mouse tissue has suggested the presence of a large number (40–70,000) of novel exons (44). Although our study identified fewer novel exons, both studies argue for the presence of novel exons in human/mouse genomes and that such novel exons can be identified using exon expression profiling.

In summary, our results indicate that exon level expression profiling can be used to molecularly classify brain tumor subgroups, can identify differentially regulated splice variants, and can identify novel exons. The splice variants identified by exon level expression profiling may lead to the identification of causative genetic changes in glial brain tumors. Furthermore, glioma-subgroup specific splice variants may serve as novel treatment targets.

Acknowledgments

Received 8/2/2006; revised 3/9/2007; accepted 3/26/2007.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

References

- Legler JM, Ries LA, Smith MA, et al. Cancer surveillance series [corrected]: brain and other central nervous system cancers: recent trends in incidence and mortality. *J Natl Cancer Inst* 1999;91:1382–90.
- Macdonald DR. New frontiers in the treatment of malignant glioma. *Semin Oncol* 2003;30:72–6.
- Kleihues P, Cavenee WK. World Health Organization Classification of Tumours of the Nervous System. Lyon: WHO/IARC; 2000.
- Hegi ME, Diserens AC, Gorlia T, et al. MGMT gene silencing and benefit from temozolomide in glioblastoma. *N Engl J Med* 2005;352:997–1003.
- Ohgaki H, Kleihues P. Epidemiology and etiology of gliomas. *Acta Neuropathol Berl* 2005;109:93–108.
- Grabowski PJ, Black DL. Alternative RNA splicing in the nervous system. *Prog Neurobiol* 2001;65:289–308.
- Nagane M, Lin H, Cavenee WK, Huang HJ. Aberrant receptor signaling in human malignant gliomas: mechanisms and therapeutic implications. *Cancer Lett* 2001; 162 Suppl:S17–21.
- Holland EC, Hively WP, DePinho RA, Varmus HE. A constitutively active epidermal growth factor receptor cooperates with disruption of G₁ cell-cycle arrest pathways to induce glioma-like lesions in mice. *Genes Dev* 1998;12:3675–85.
- Mellinghoff IK, Wang MY, Vivanco I, et al. Molecular determinants of the response of glioblastomas to EGFR kinase inhibitors. *N Engl J Med* 2005;353: 2012–24.
- Frederick L, Wang XY, Eley G, James CD. Diversity and frequency of epidermal growth factor receptor mutations in human glioblastomas. *Cancer Res* 2000;60: 1383–7.
- Shi E, Kan M, Xu J, Wang F, Hou J, McKeenan WL. Control of fibroblast growth factor receptor kinase signal transduction by heterodimerization of combinatorial splice variants. *Mol Cell Biol* 1993;13:3907–18.
- Wang F, Kan M, Yan G, Xu J, McKeenan WL. Alternately spliced NH₂-terminal immunoglobulin-like Loop I in the ectodomain of the fibroblast growth factor (FGF) receptor 1 lowers affinity for both heparin and FGF-1. *J Biol Chem* 1995;270:10231–5.
- Chunduru S, Kawami H, Gullick R, Monacci WJ, Dougherty G, Cutler ML. Identification of an alternatively spliced RNA for the Ras suppressor RSU-1 in human gliomas. *J Neurooncol* 2002;60:201–11.
- Simon M, Koster G, Ludwig M, et al. Alternative splicing of the p15 cdk inhibitor in glioblastoma multiforme. *Acta Neuropathol Berl* 2001;102:167–74.
- Uno M, Oba-Shinjo SM, de Aguiar PH, et al. Detection of somatic TP53 splice site mutations in diffuse astrocytomas. *Cancer Lett* 2005;224:321–7.
- Xu X, Joh HD, Pin S, et al. Expression of multiple larger-sized transcripts for several genes in oligodendrogliomas: potential markers for glioma subtype. *Cancer Lett* 2001;171:67–77.
- Hui L, Zhang X, Wu X, et al. Identification of alternatively spliced mRNA variants related to cancers by genome-wide ESTs alignment. *Oncogene* 2004;23: 3013–23.
- French PJ, Swagemakers SMA, Nagel JHA, et al. Gene expression profiles associated with treatment response in oligodendrogliomas. *Cancer Res* 2005;65: 11335–44.
- Schroeder A, Mueller O, Stocker S, et al. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol Biol* 2006;7:3.
- Valk PJ, Verhaak RG, Beijin MA, et al. Prognostically useful gene-expression profiles in acute myeloid leukemia. *N Engl J Med* 2004;350:1617–28.
- Mischel PS, Cloughesy TF, Nelson SF. DNA-microarray analysis of brain cancer: molecular classification for therapy. *Nat Rev Neurosci* 2004;5:782–92.
- Rickman DS, Bobek MP, Misek DE, et al. Distinctive molecular profiles of high-grade and low-grade gliomas based on oligonucleotide microarray analysis. *Cancer Res* 2001;61:6885–91.
- Godard S, Getz G, Delorenzi M, et al. Classification of human astrocytic gliomas on the basis of gene

- expression: a correlated group of genes with angiogenic activity emerges as a strong predictor of subtypes. *Cancer Res* 2003;63:6613–25.
24. Nutt CL, Mani DR, Betensky RA, et al. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res* 2003;63:1602–7.
25. Freije WA, Castro-Vargas FE, Fang Z, et al. Gene expression profiling of gliomas strongly predicts survival. *Cancer Res* 2004;64:6503–10.
26. Liang Y, Diehn M, Watson N, et al. Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme. *Proc Natl Acad Sci U S A* 2005;102:5814–9.
27. Phillips HS, Kharbanda S, Chen R, et al. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell* 2006;9:157–73.
28. Ule J, Ule A, Spencer J, et al. Nova regulates brain-specific splicing to shape the synapse. *Nat Genet* 2005; 37:844–52.
29. Xu Q, Modrek B, Lee C. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res* 2002;30:3754–66.
30. Yeo G, Holste D, Kreiman G, Burge CB. Variation in alternative splicing across human tissues. *Genome Biol* 2004;5:R74.
31. Pan Q, Shai O, Misquitta C, et al. Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Mol Cell* 2004; 16:929–41.
32. Black DL. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* 2003;72:291–336.
33. Kalnina Z, Zayakin P, Silina K, Line A. Alterations of pre-mRNA splicing in cancer. *Genes Chromosomes Cancer* 2005;42:342–57.
34. Graveley BR. Sorting out the complexity of SR protein functions. *RNA* 2000;6:1197–211.
35. Frederick L, Eley G, Wang XY, James CD. Analysis of genomic rearrangements associated with EGRFvIII expression suggests involvement of Alu repeat elements. *Neuro-oncol* 2000;2:159–63.
36. Johnson JM, Castle J, Garrett-Engle P, et al. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* 2003; 302:2141–4.
37. Yeakley JM, Fan JB, Doucet D, et al. Profiling alternative splicing on fiber-optic arrays. *Nat Biotechnol* 2002;20:353–8.
38. Zhu J, Shendure J, Mitra RD, Church GM. Single molecule profiling of alternative pre-mRNA splicing. *Science* 2003;301:836–8.
39. Fehlbauer P, Guihal C, Bracco L, Cochet O. A microarray configuration to quantify expression levels and relative abundance of splice variants. *Nucleic Acids Res* 2005;33:e47.
40. Nagao K, Togawa N, Fujii K, et al. Detecting tissue-specific alternative splicing and disease-associated aberrant splicing of the PTCH gene with exon junction microarrays. *Hum Mol Genet* 2005;14:3379–88.
41. O'Connor V, Genin A, Davis S, et al. Differential amplification of intron-containing transcripts reveals long term potentiation-associated up-regulation of specific Pde10A phosphodiesterase splice variants. *J Biol Chem* 2004;279:15841–9.
42. Chen FC, Chen CJ, Ho JY, Chuang TJ. Identification and evolutionary analysis of novel exons and alternative splicing events using cross-species EST-to-genome comparisons in human, mouse and rat. *BMC Bioinformatics* 2006;7:136.
43. Nekrutenko A. Identification of novel exons from rat-mouse comparisons. *J Mol Evol* 2004;59:703–8.
44. Frey BJ, Mohammad N, Morris QD, et al. Genome-wide analysis of mouse transcripts using exon microarrays and factor graphs. *Nat Genet* 2005;37:991–6.

Cancer Research

The Journal of Cancer Research (1916–1930) | The American Journal of Cancer (1931–1940)

Identification of Differentially Regulated Splice Variants and Novel Exons in Glial Brain Tumors Using Exon Expression Arrays

Pim J. French, Justine Peeters, Sebastiaan Horsman, et al.

Cancer Res 2007;67:5635-5642.

Updated version	Access the most recent version of this article at: http://cancerres.aacrjournals.org/content/67/12/5635
Supplementary Material	Access the most recent supplemental material at: http://cancerres.aacrjournals.org/content/suppl/2007/06/12/67.12.5635.DC1

Cited articles	This article cites 43 articles, 13 of which you can access for free at: http://cancerres.aacrjournals.org/content/67/12/5635.full#ref-list-1
Citing articles	This article has been cited by 10 HighWire-hosted articles. Access the articles at: http://cancerres.aacrjournals.org/content/67/12/5635.full#related-urls

E-mail alerts	Sign up to receive free email-alerts related to this article or journal.
Reprints and Subscriptions	To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org .
Permissions	To request permission to re-use all or part of this article, use this link http://cancerres.aacrjournals.org/content/67/12/5635 . Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site.