Abstract. Recent advents in magnetic resonance spectroscopy (MRS) techniques permit subsequent microarray analysis over the entire human transcriptome in the same tissue biopsies. However, extracting information from such immense quantities of data is limited by difficulties in recognizing and evaluating the relevant patterns of apparent gene expression in the context of the existing knowledge of phenotypes by histopathology. Using a quantitative approach derived from a knowledge base of pathology findings, we present a novel methodology used to process genome-wide transcription and MRS data. This methodology was tested to examine metabolite and genome-wide profiles in MRS and RNA in 55 biopsies from human subjects with brain tumors with ~100% certainty. With the guidance of histopathology and clinical outcome, 15 genes with the assistance of 15 MRS metabolites were able to be distinguished by tumor categories and the prediction of survival was better than when either method was used alone.

Introduction

According to the Central Brain Tumor Registry of the USA (www.cbtrus.org), the worldwide incidence rate of primary malignant brain and central nervous system (CNS) tumors, age-adjusted using the world standard population, is 3.7 per 100,000 person-years in males and 2.6 per 100,000 person-years in females. The incidence rates are higher in more developed countries (males, 5.8 per 100,000 person-years; females, 4.1 per 100,000 person-years) than in less developed countries (males, 3.0 per 100,000 person-years; females, 2.1 per 100,000 person-years).

Management of brain tumors in patients would benefit from improved characterization, diagnosis and prognostic biomarkers. The diagnostic utility of biomarkers for tissue typing lies in their biological relevance. Highly informative biomarker profiles are difficult to establish, due to current technical limitations and the small sample sizes of tissue biopsies, which pose challenges for producing accurate magnetic resonance spectroscopy (MRS) and transcriptome data. Further development and application of microscale MRS and genomics can overcome these limitations and identify new biomarkers to accurately type cancers. This report focuses on applying these technologies to brain tumors, the leading cause for high mortality in older adults (1,2).
Our hypothesis is that current tissue characterization is enhanced by developing and applying a classification strategy analysis algorithm that produces unique tumor fingerprints by combining biomarker profiles from ex vivo MRS and whole-genome expression profiling performed on microscale pediatric brain tumor biopsies. Brain-tumor specific biomarkers can be identified using high-resolution ex vivo MRS at high magnetic field strengths and a combination of microarray, bioinformatics, and computational analyses. In the near future, combining clinical MRI, MRS and MR imaging of gene expression in vivo should produce superior images to enhance the specificity of cancer diagnosis in clinical medicine. Knowledge acquired from these studies can also be used to type inoperable cancers. This hypothesis builds upon prior reports (3-16). Prior data have provided the following information and advances: a) measurements using one-dimensional (1D) ex vivo High-Resolution Magic Angle Spinning (HRMAS) proton (1H) MRS at 9.4 T correlated directly with the neuropathology of intact brain tumor biopsies. The intracellular metabolite phosphocholine (PCho), a constituent of the choline (Cho) peak, was associated with cellularity and proliferative activity (17); b) MRS-detectable lipid changes were implicated in brain tumor apoptosis and necrosis (17); c) in vivo MRS spectra correlated with ex vivo MRS measurements (17); d) Gene expression analyses of embryonal CNS tumors distinguished between tumor types that could not be clearly distinguished by histopathology (18,19); e) HRMAS 1H MRS and genomic analyses of microscale tissue biopsies (≥2 mg) yielded quality data, enabling MRS-derived metabolites and gene expression differences to be related and cross validated (15); f) an optimized adiabatic solid-state NMR method, Total Through-Bond Spectroscopy (TOBSY), maximized the advantages of HRMAS applied to intact biopsies when compared to more conventional liquid-state NMR approaches (20,21); g) a structured network knowledge-based approach was demonstrated to be capable of analyzing genome-wide transcriptional responses in the context of known functional interrelationships among proteins, small molecules, and phenotypes (13); and h) the design and initial testing of our classification algorithms was successful (22). While considerable work has been done on the classification of cancers based on genomic data (23-33), and some work has been done using MRS data (22,34-42), these two datasets have yet to be integrated along with other clinical features. We hope to improve cancer diagnosis accuracy immediately following biopsy collection by uncovering and exploiting complementary information in the MRS and genomic data. We undertook this challenge by employing machine-learning methods that combine the data generated from MRS and genomics. Our rationale for using methods that emerged under the machine-learning framework, such as the support vector machines (SVM), is that these methods have already been employed successfully for cancer classification for more than ten years (25,32,43-45) and have replaced other traditional methods such as linear discriminant analysis, logistic regression, density-based methods (Parzen window, Naïve Bayes), neural networks with PCA pre-processing, and decision trees (32). As a result of its wide acceptance as a state-of-the-art method for gene-based diagnosis, the SVM constitutes an important module of modern software tools for gene expression analysis (46). On the other hand, statistical methods are limited mostly to the task of ranking individual genes with respect to their ability to act as individual markers, thus they can be useful in the feature selection task. Also, the SVM approach has been proposed for classification of heterogeneous data sources using a weighted combination of multiple similarity measures (kernels), (one kernel for each data source). This methodology has already been successfully applied in several medical diagnosis tasks (47,48) and is a promising technique for combining data from multiple sources. In our line of research, the data sources to be combined are genomics, MRS, and clinical data. Here, we report that a new classification system using the sensor fusion approach, which combines genomics, MRS, statistics and biological content, improves the typing and understanding of the complexity of human brain tumors, as well as the search for novel tumor biomarkers, an important step for novel drug development. It also generates testable hypotheses regarding neoplasia and promises to guide human brain tumor therapy provided that improved in vivo methods for monitoring response to therapy are developed.

Materials and methods

Experimental design. We carried out experiments on a dataset of 55 gene expression profiles derived from normal (9 cases) and tumor (46 cases) classes. The tumor class samples belonged to three categories: high grade (H) [20 cases, 12 glioblastoma multiforme (GBM); 8 anaplastic astrocytoma (AA)], low grade (L) (17 cases, 7 meningioma; 7 schwannoma; 7 pylocytic astrocytoma) and metastasized (M) (11 cases, 5 adenocarcinoma; 3 breast cancer metastasis; 3 other metastasis). Subjects ranged in age from 17 to 54 years.

HRMAS 1H MRS using adiabatic TOBSY. We used a previously designed 2D ex vivo HRMAS 1H MRS procedure for brain tumors, based on novel concepts rooted in solid-state NMR spectroscopy (49). All HRMAS 1H MRS using TOBSY experiments were performed on a Bruker BioSpin Avance NMR spectrometer (600.13 MHz) using a 4-mm triple resonance (1H, 13C, 1H) HRMAS probe (Bruker). Specimens were pre-weighted and transferred to a ZrO2 rotor tube (4 mm diameter, 50 μl), containing an external standard [trimethylsilyl propionic-2,2,3,3-d4 acid (TSP), Mw=172, δ=0.00 ppm] that functioned as a reference both for resonance chemical shift and quantification. The HRMAS 1H MRS was performed at -8°C with 3 kHz MAS to minimize tissue degradation. Typical acquisition parameters were, 2 k points direct dimension (13 ppm spectral width), 200 points indirect dimension (7.5 ppm spectral width), 8 scans with 2 dummy scans, 1 sec water pre-saturation, 2 sec total repetition time, 45 msec mixing time and total acquisition time 45 min.

Analysis of 2D TOBSY MR spectra. The spectra of intact specimens were analyzed using the XWINNMR 3.5 software package (Bruker Biospin Corp, Billerica, MA). Before Fourier transformation and phasing, the 2D free induction decays were subjected to QINS=3 window apodization. Baseline correction was then performed using a low order spline function. After Lorentzian and/or Gaussian fitting, the area under the curves or the volumes of the 15 most intense spectra resonances were
calculated. Relative quantification using the TSP standard was performed as described below. These resonances were identified and assigned to the corresponding metabolites.

**Quantification of brain metabolites from the 2D TOBY MR spectra.** To quantify the brain metabolites, we used the ratio of the cross peak volumes of the metabolites \[ \text{CPV}(M) \] to the TSP diagonal peak volume \[ \text{DPV}(\text{TSP}) \]. This ratio was further divided by the biopsy weight \( w \) to yield the normalized metabolite intensity, \( I_c = \frac{1}{w} \times \frac{\text{CPV}(M)}{\text{DPV}(\text{TSP})} \).

Microscale transcriptome analyses to determine the gene expression profiles of the tumor biopsy samples after HRMAS \(^1H\) MRS. We performed microscale genome array studies with the commercially available Affymetrix U133Plus® array (Santa Clara, CA).

**Platform choice.** The Affymetrix GeneChip® DNA microarray platform has several significant advantages over competing technologies, including coverage of the entire human genome, access to probe sequences, probe redundancy (11 sequences per gene) to optimize fidelity of the signal-to-noise ratio, ready commercial availability, standardization of hybridization, washing, staining and scanning processes, quality control built into the manufacturing processes, available technical support, and a relatively low cost per investigated gene.

**RNA purification.** Total experimental RNA were isolated from the biopsy samples used for HRMAS \(^1H\) MRS. Total control RNA was isolated from normal tissue removed along with the tumor biopsies, or from age-matched patients undergoing epilepsy surgery. RNA was isolated using the modified protocol of the RNeasy purification kit (Qiagen) that our Sanford colleagues have optimized. Briefly, during tissue homogenation and deproteination, 1 mg of tRNA and 10 mg polyacrylamide are added as carriers. This greatly improves RNA yields to ~500 ng of total RNA per mg of tissue, an amount x20 greater than that required for our optimized RNA labeling procedure. RNA purity was assessed from the OD 260/280 ratio, with only samples having ratios >1.9 retained for further use. In addition, RNA integrity was assessed by the Agilent 2100 Bioanalyzer, where good quality samples exhibit a relatively flat and low baseline in the capillary electrophoresis elution and have 18S and 28S peaks between 1:1 to 1:2, as scored by the Bioanalyzer software.

**RNA labeling.** We used the Ribo-SPIA protocol (www.nugeninc.com) for mRNA labeling and amplification. Ribo-SPIA is superior to all other labeling methods, when the amount of RNA is <1 mg. An overview of the Ribo-SPIA amplification process was described previously by Tzika et al., (15). We used 20 ng total RNA for first strand cDNA synthesis, and the entire procedure for amplification, fragmentation and labeling was performed in one day.

**Data analysis of gene expression.** The expression profiles of tumor biopsies analyzed and compared to those of control tissue. Specifically, the raw Affymetrix CEL files were normalized and analyzed to obtain expression values using both dChip (http://biosun1.harvard.edu/~tibs/dchip/) and GC-RMA (50,51). Both sets of obtained expression values analyzed (as described below) and the results from the two methods were compared. We used significant analysis of microarrays (SAMs) (http://www-stat.stanford.edu/~tibs/SAM/) to obtain a list of differentially expressed genes with a false discovery rate (q-value) <0.05 and to properly take into account the substantial multiple comparison problem (52,53). BRB-Array Tools (http://linus.nci.nih.gov/BRB-ArrayTools.html) were used to classify tumor types based on expression patterns, using several major built-in methods including SVM and Bayesian classifier. BRB-Array software was used to divide samples into training and test datasets and to perform gene selection and model building in each cross-validation run of the training dataset. This cross-validation ensured an unbiased final prediction accuracy for the test dataset. The top genes selected by the classifiers were clustered and visualized in dChip. The enriched Gene Ontology and pathway groups in these top genes were identified by dChip and Ingenuity Pathway Analysis software and correlated with pathways implicated in CNS tumorigenesis. Analysis results were produced in both tabular and graphical formats.

**Classification strategy.** The architecture of our classification system is shown in Fig. 1. We first fine-tuned the feature selection process by which the high dimensionality of the SMAs output was reduced by selecting only the most relevant genes for the classification task. Then, a classifier was constructed to these reduced feature vectors in order to optimally partition the space according to class. We chose to use the SVM classifier (54). Finally, the constructed reduced feature space from the gene expression values was combined with the NMR features in order to examine their impact on the classifier.

**Feature selection (FS).** Feature selection methods typically rank genes according to their differential expressions among phenotypes and pick the top-ranked genes. There are two general schemes for feature selection, filters and wrappers (50,55). We used the minimum redundancy - maximum relevance (MRMR) method (51), because it is a powerful framework for selecting features that capture class characteristics in a broad spectrum by reducing mutual redundancy within the feature set. Thus, it offers greater robustness and generalization properties to the reducing feature space of samples, which can significantly improve classification accuracy.
SVM classifier. SVM (54) is a very powerful classification method that draws hyperplanes in the feature vector space by maximizing the margin between data samples of different classes. SVM is built upon the use of kernels to construct nonlinear decision boundaries. Here, we used linear kernels and the LIBSVM environment for multi-class SVMs (52). It should be noted that during all experiments with SVM, we adopted the standard leave-one-out training/testing scheme. That is, one element of the data was used as a training set, and the left-out element was used for testing the predictive performance of the resulting classifier. The SVM soft-margin constant C was set to 10, chosen based on the results of a few runs on one training set. The results indicated that the value of this parameter was not crucial for our experimental dataset.

Statistical analysis. Multiple stepwise logistic regression analysis was done to evaluate whether genomic and/or HRMAS MRS data can predict clinical outcome. Maximum likelihood estimation of the logistic model provided coefficients, SEs, adjusted odds ratios, 95% confidence intervals, the likelihood ratio $\chi^2$ test for parameters as well as sensitivity, specificity and accuracy of the prediction of the clinical outcome. Statistical analysis was conducted with the SPSS software package (version 16.0, SPSS Inc., Chicago, IL), and two-tailed P-values of <0.05 were considered statistically significant.

Results

The impact of each the following 16 NMR features on the classifier were examined, alanine (Ala), aspartate (Asp), choline (Cho), ethanolamine (Etn), $\gamma$-aminobutyric acid (GABA), glutamine (Gln), glutamate (Glu), glycerophosphocholine (GPC), lactate (Lac), lipids (Lip), myoinositol (Myo), N-acetyl aspartate (NAA), phosphocholine (PC), phospho-ethanolamine (PE), polyunsaturated fatty acids (PUFA) and taurine (Tau). Typical 2D TOBSY MR spectra are shown in Fig. 2. A classification strategy analysis algorithm to identify combinatorial biomarker profiles that uniquely define tumor

Statistical analysis. Multiple stepwise logistic regression analysis was done to evaluate whether genomic and/or HRMAS MRS data can predict clinical outcome. Maximum likelihood estimation of the logistic model provided coefficients, SEs, adjusted odds ratios, 95% confidence intervals, the likelihood ratio $\chi^2$ test for parameters as well as sensitivity, specificity and accuracy of the prediction of the clinical outcome. Statistical analysis was conducted with the SPSS software package (version 16.0, SPSS Inc., Chicago, IL), and two-tailed P-values of <0.05 were considered statistically significant.

Results

The impact of each the following 16 NMR features on the classifier were examined, alanine (Ala), aspartate (Asp), choline (Cho), ethanolamine (Etn), $\gamma$-aminobutyric acid (GABA), glutamine (Gln), glutamate (Glu), glycerophosphocholine (GPC), lactate (Lac), lipids (Lip), myoinositol (Myo), N-acetyl aspartate (NAA), phosphocholine (PC), phospho-ethanolamine (PE), polyunsaturated fatty acids (PUFA) and taurine (Tau). Typical 2D TOBSY MR spectra are shown in Fig. 2. A classification strategy analysis algorithm to identify combinatorial biomarker profiles that uniquely define tumor

Figure 2. Typical TOtal Through-Bond SpectroscopY (TOBSY) using ex vivo HRMAS MRS on anaplastic astrocytoma (left) and meningioma (right) biopsies. HRMAS $^1$H MR spectra using TOBSY, with 45 msec mixing time, 3 kHz MAS speed, and -8°C at 600 MHz. (Ala, alanine; Cho, choline; GABA, $\gamma$-aminobutyric acid; Gln, glutamine; Glu, glutamate; GPC, glycerophosphocholine; Lip, lipids; Myo, myoinositol; PC, phosphocholine; PE, phospho-ethanolamine; PUFA, polyunsaturated fatty acids; Tau, taurine). Note that Cho, PC, GPC, PE, Etn, are clearly separable here due to the use of the 2D TOBSY method. Also note that the anaplastic astrocytoma (high-grade) exhibits different MR spetrum as compared to meningioma (low-grade).

Figure 3. The rich feature space created by the two first selected genes (shown in x and y axis) that is linearly divided into normal and tumor class regions; 202126_at is PRP4-pre-mRNA processing factor 4 (gene symbol, PRPF4B); 202508_s_at is synaptosomal associated protein, 25 kDa (gene symbol SNAP25). Further details on the biological function(s) of these genes and correlation with diseases are included in Table I.
stem cell marker for malignant brain tumors, cd133 (53); iii) tumor typing and subtype classification. As expected, this classification scheme was more difficult. The classifier had near excellent behavior using the first 9-25 features selected by the MRMR method (Table I). Using the first 15 genes, the classifier reached 100% best accuracy. Specifically, after adding gene 209771_x_at to the feature vectors, the classification performance was 80.4%; adding gene 229851_s_at increased the performance to 91.3%; adding gene 225491_at increased it to 95.6%; adding genes 211991_s_at, 1552797_s_at, 224209_s_at, 206349_at, 204131_s_at and 241938_at increased it to 97.8%. Finally, the addition of gene 204501_s_at increased the classification performance to 100% for anaplastic astrocytoma and meningioma. We also tested the impact of combining selected NMR features (isolated or combined) with gene values. Using all 15 NMR features and selected genes (obtained from the MRMR genes selection method) increased the classification performance from 95.6% to 97.8% for the high grade typing (11 genes and 15 NMR features), from 95.6% to 100% for schwannoma subtyping (1 gene 209169_at and 15 NMR features) and from 95.6% to 97.8% for metastasis subtyping (12 genes and 15 NMR features).

Studies demonstrating the potential of HRMAS combined with gene expression profiles offer better accuracy than each methodology alone in predicting survival. We performed multiple stepwise logistic regression analysis to evaluate how gene expression values, HRMAS MRS data, and their combination predict survival. We chose the 15 best genes according to their MRMR algorithm rank (Table I) and 15 metabolite values (Ala, Asp, Cho, Etn, GABA, Gln, Glu, GPC, Lac, Lip, Myo, NAA, PC, PE, PUFA, Tau), (see above) corresponding to 49 available binary clinical outcomes (33 survived vs. 16 deceased). Our preliminary results have proven that the combination of genomic and HRMAS MRS data improves the ability to predict clinical outcome. More specifically, gene data alone achieved high sensitivity, predicting 15 out of 16 deceased cases (sensitivity 94%), high specificity, predicting 32 out of 33 cases (specificity 97%), and high accuracy (96%). HRMAS MRS data had inferior sensitivity (11/16, 69%), specificity (28/32, 85%), and accuracy (80%). Combining genomics and HRMAS MRS data logistic regression achieved a perfect classification (100% for all indices) of survived and deceased cases. Although we believe that these promising results are affected by sample size, they clearly demonstrate that the combination of gene expression and MRS data predict a clinically meaningful parameter, such as survival, better than either technique alone.

Discussion

Our objective in this study was to use a novel approach that combines biomarkers detected with magnetic resonance spectroscopy (MRS) and molecular genomics to improve the characterization and prognostication of biospecimens in molecular medicine. We aimed to develop a useful clinical tool that uses tissue fingerprinting to aid clinicians not only in making diagnostic and treatment course decisions, but also in understanding the biology of brain malignancy subtypes in humans, an important step for novel drug development. The method is based on the development of a classification strategy analysis algorithm that combines biomarker profiles generated using high-resolution ex vivo MRS and whole-genome expression analyses of microscale brain tumor samples as well as features from the clinical patient database (i.e., survival). We optimized and applied ex vivo HRMAS 1H MRS and transcriptome profiling to intact tumor biopsies that are <2 mg. We then combined these data sets to develop a classification strategy analysis algorithm to produce tissue fingerprints that accurately type these biopsies and demonstrated the potential of HRMAS and its combination with gene expression profiles to offer better accuracy than each methodology alone in predicting survival. Herein, we demonstrated for the first time that a combined approach of using metabolite and gene expression profiles allows for more accurate discernment of tumor categories and better prediction of patient survival than either method alone. Previous data have shown that metabolites derived from brain proton MRS predict clinically meaningful parameters such as treatment response and survival of children with CNS tumors (10,56,57). Gene expression has been reported to predict outcome and survival with greater accuracy than histology (18,19). Prior studies have demonstrated that ex vivo MRS can be used to classify brain tumors with high sensitivity, specificity, and accuracy using only 16 metabolites, which is the highest number of metabolites detected with in vivo 2D approaches (22). Importantly, the agreement between ex vivo and in vivo MRS data suggests that in vivo 2D MRS data agree (17). It was thus important to train our algorithm (which can handle missing data) with multiple kinds of information, including both MRS and gene expression data, to increase its discriminatory capability that will also allow its application in vivo and specifically in cases of inoperable tumors. Although the ability to produce tissue MRS data non-invasively using 2D in vivo MRS would provide a considerable advantage, it is beyond the scope of the current investigation. In vivo MRS is clinically feasible and we have used it previously to investigate whether ex vivo and in vivo MRS data agree (17). We found that such agreement exists and thus obviates the use of in vivo MRS to guide ex vivo biopsy collection for
<table>
<thead>
<tr>
<th>Probe Set ID</th>
<th>Gene symbol</th>
<th>Gene title</th>
<th>Biological process term</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>1552797_s_at</td>
<td>PROM2</td>
<td>prominin 2</td>
<td>Iron ion binding, electron carrier activity, transmembrane receptor activity, growth factor activity, role in cell death</td>
<td>Differential expression between chromophobe RCC and oncocytoma (66)</td>
</tr>
<tr>
<td>200853_at</td>
<td>H2AFZ</td>
<td>H2A histone family, member Z</td>
<td>Nucleosome assembly, chromosome organization and biogenesis</td>
<td>Cardiac myocyte hypertrophy (67)</td>
</tr>
<tr>
<td>201839_s_at</td>
<td>TACSTD1</td>
<td>Tumor-associated calcium signal transducer 1; aka EpCAM</td>
<td>Gene encodes protein for plasma membrane, integral to membrane, extracellular space, role in cell metabolism, colony formation and cell proliferation</td>
<td>Ovarian (68), breast (69), esophageal (70), prostate cancer metastasis</td>
</tr>
<tr>
<td>202126_at</td>
<td>PRPF4B</td>
<td>PRP4 pre-mRNA processing factor 4 homolog B (yeast)</td>
<td>mRNA processing, protein amino acid phosphorylation, RNA splicing</td>
<td>Diseases involving PRP4-mediated phosphorylation of KLF13 plays a role in the regulation of CCL5 expression in T lymphocytes (71)</td>
</tr>
<tr>
<td>202286_s_at</td>
<td>TACSTD2</td>
<td>Tumor-associated calcium signal transducer 2</td>
<td>Regulation of progression through cell cycle, cell surface receptor linked signal transduction, cell proliferation, visual perception, role in cell proliferation expressed in most human carcinomas</td>
<td>Gelatin-lattice corneal dystrophy type III (72,73)</td>
</tr>
<tr>
<td>202508_s_at</td>
<td>SNAP25</td>
<td>Synaptosomal-associated protein, 25 kDa</td>
<td>Neurotransmitter uptake, synaptic transmission, neurotransmitter secretion, synaptic vesicle docking during exocytosis, regulation of insulin secretion, exocytosis, axonogenesis, long-term memory, endosome transport, growth hormone secretion, sleep, positive regulation of insulin secretion, positive regulation of hormone secretion, calcium ion-dependent exocytosis of neurotransmitter, regulation of synaptogenesis, role in cell fusion, exocytosis and release</td>
<td>Schizophrenia (74), pituitary adenomas (75)</td>
</tr>
<tr>
<td>202748_at</td>
<td>GBP2</td>
<td>Guanylate binding protein 2, interferon-inducible</td>
<td>Immune response, role in cell contact growth inhibition</td>
<td>Esophageal squamous cell carcinomas (76)</td>
</tr>
<tr>
<td>204030_s_at</td>
<td>SCHIP1</td>
<td>Schwannomin interacting protein 1</td>
<td>Protein binding, protein homodimerization activity, identical protein binding</td>
<td>No data available</td>
</tr>
<tr>
<td>204501_at</td>
<td>NOV</td>
<td>Nephroblastoma overexpressed gene</td>
<td>Regulation of cell growth, role in cell migration, growth and invasion</td>
<td>Adrenocortical tumors, astrocytomas, multiple sclerosis (77), metastatic melanoma (78), renal cell carcinoma (79)</td>
</tr>
<tr>
<td>204587_at</td>
<td>SLC25A14</td>
<td>Solute carrier family 25 (mitochondrial carrier, brain), member 14</td>
<td>Transport, aerobic respiration, mitochondrial transport, role in cell transmembrane potential and mitochondrial uncoupling</td>
<td>Amplified in brain ischemia, schizophrenia (80-82)</td>
</tr>
<tr>
<td>205933_at</td>
<td>SETBP1</td>
<td>SET binding protein 1</td>
<td>Regulation of transcription, DNA-dependent</td>
<td>Treatment of X-linked chronic granulomatous disease (X-CGD) (83)</td>
</tr>
</tbody>
</table>
Table I. Continued.

<table>
<thead>
<tr>
<th>Probe set ID</th>
<th>Gene symbol</th>
<th>Gene title</th>
<th>Biological process term</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>206018_at</td>
<td>FOXG1</td>
<td>Forkhead box G1</td>
<td>Transcription, regulation of transcription, DNA-dependent, multicellular organismal development, brain development, positive regulation of neuroblast proliferation, regulation of mitotic cell cycle, dorsal/ventral pattern formation, axon midline choice point recognition, central nervous system neuron development, forebrain development, inner ear morphogenesis, negative regulation of neuron differentiation, positive regulation of cell cycle, neuron morphogenesis during differentiation, role in cell differentiation, morphogenesis and growth</td>
<td>Bladder cancer (84), Rett syndrome (85), medulloblastoma (86)</td>
</tr>
<tr>
<td>208998_at</td>
<td>UCP2</td>
<td>Uncoupling protein 2 (mitochondrial, proton carrier)</td>
<td>Transport, mitochondrial transport, response to superoxide, proton transport, role in cell transmembrane potential, cell respiration and cell quantity</td>
<td>Obesity, diabetes (87,88)</td>
</tr>
<tr>
<td>209496_at</td>
<td>RARRES2</td>
<td>Retinoic acid receptor responder (tazarotene induced) 2</td>
<td>Retinoid metabolic process, defense response role in cell migration, transmigration and differentiation</td>
<td>Adrenocortical carcinoma (89), ovarian endometriosis (90), APC-mutant intestinal adenomas (91), Insulin resistance (92), insulin resistance in polycystic ovary syndrome (93), obesity and lipodystrophy (94)</td>
</tr>
<tr>
<td>212274_at</td>
<td>LPIN1</td>
<td>Lipin 1</td>
<td>Lipid metabolic process, organ regeneration, ruffle organization and biogenesis, actin cytoskeleton reorganization, cellular response to insulin stimulus, regulation of fat cell differentiation, role in cell differentiation, enlargement and appearance</td>
<td>Neuroblastom development and survival prediction (95), Alzheimer's disease pathogenicity (96), breast carcinomas outcome prediction (97), prostate cancer (98), P. aeruginosa infection (99), metastasis (100,101), catalepsy (102), kindling (103)</td>
</tr>
<tr>
<td>212486_s_at</td>
<td>FYN</td>
<td>FYN oncogene related to SRC, FGR, YES</td>
<td>Protein amino acid phosphorylation, calcium ion transport, protein kinase cascade, multicellular organismal development, learning, feeding behavior, interspecies interaction between organisms, T cell receptor signaling pathway, neuron migration, cell surface receptor linked signal transduction, regulation of cell shape, peptidyl-tyrosine phosphorylation, forebrain development, myelination, response to ethanol, protein amino acid autophosphorylation, activated T cell proliferation, detection of mechanical stimulus involved in sensory perception of pain, ionotropic glutamate receptor signaling pathway, role in cell proliferation, quantity and activation</td>
<td>Neuroblastoma development and survival prediction (95), Alzheimer's disease pathogenicity (96), breast carcinomas outcome prediction (97), prostate cancer (98), P. aeruginosa infection (99), metastasis (100,101), catalepsy (102), kindling (103)</td>
</tr>
<tr>
<td>212565_at</td>
<td>STK38L</td>
<td>Serine/threonine kinase 38 like</td>
<td>Protein amino acid phosphorylation, protein kinase cascade, regulation of cellular component organization and biogenesis, protein targeting, role in cell outgrowth and invasion</td>
<td>No data available</td>
</tr>
</tbody>
</table>
Table I. Continued.

<table>
<thead>
<tr>
<th>Probe set ID</th>
<th>Gene symbol</th>
<th>Gene title</th>
<th>Biological process term</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>212632_at</td>
<td>STX7</td>
<td>Syntaxin 7</td>
<td>Intracellular protein transport, post-Golgi vesicle-mediated transport, vacuole organization and biogenesis, vesicle-mediated transport, synaptic vesicle exocytosis, role in cell fusion and transport</td>
<td>Melanoma (104), adenomyosis (105)</td>
</tr>
<tr>
<td>212708_at</td>
<td>MSL-1</td>
<td>Male-specific lethal-1 homolog</td>
<td>A multisubunit human histone acetylase complex responsible for histone H4 lysine-16 acetylation of all cellular chromosomes</td>
<td>No data available</td>
</tr>
<tr>
<td>213693_s_at</td>
<td>MUC1</td>
<td>Mucin 1, cell surface associated</td>
<td>This gene encodes a membrane bound, glycosylated phosphoprotein, which serves a protective function by binding to pathogens and also functions in a cell signaling capacity, role in cell apoptosis, growth and lysis</td>
<td>Breast, prostate, colon, stomach, lung, ovary and pancreas adenocarcinomas (106), intrahepatic cholangiocarcinoma metastasis and outcome (107)</td>
</tr>
<tr>
<td>214512_s_at</td>
<td>SUB1</td>
<td>SUB1 homolog (S. cerevisiae)</td>
<td>Transcription, regulation of transcription from RNA polymerase II promoter, regulation of transcription, DNA-dependent, role in cell apoptosis</td>
<td>Involved in replication of simian virus 40 and adenov-associated virus (AAV) (108,109)</td>
</tr>
<tr>
<td>218035_s_at</td>
<td>FLJ20273</td>
<td>RBM47 - RNA binding motif protein 47</td>
<td>RNA binding, nucleotide binding, nucleic acid binding</td>
<td>No data available</td>
</tr>
<tr>
<td>219368_at</td>
<td>NAP1L2</td>
<td>Nucleosome assembly protein 1-like 2</td>
<td>Nucleosome assembly, role in cell assembly and quantity</td>
<td>Colon cancer (110)</td>
</tr>
<tr>
<td>224209_s_at</td>
<td>GDA</td>
<td>Guanine deaminase</td>
<td>Response to stimulus, nucleobase, nucleoside, nucleotide and nucleic acid metabolic process, nervous system development</td>
<td>Viral hepatitis (111), learning and memory (112), <a href="http://nootropics.com/cytoindex.html">http://nootropics.com/cytoindex.html</a>, severe shock (113), myocardial infarction (114,115)</td>
</tr>
<tr>
<td>225321_s_at</td>
<td>PILRB</td>
<td>Paired immunoglobulin-like type 2 receptor β</td>
<td>Transmembrane receptor protein tyrosine kinase signaling pathway transmembrane receptor protein tyrosine kinase activation (dimerization)</td>
<td>BCR-Able positive ALL (119), cancer (120), S. aureus-induced pneumonia (121), autoimmune disorders (122)</td>
</tr>
</tbody>
</table>
### Table I. Continued.

<table>
<thead>
<tr>
<th>Probe set ID</th>
<th>Gene symbol</th>
<th>Gene title</th>
<th>Biological process term</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>228335_at</td>
<td>CLDN11</td>
<td>Claudin 11 (oligodendrocyte transmembrane protein)</td>
<td>Cell adhesion, spermatogenesis, axon ensheathment, calcium-independent cell-cell adhesion, role in cell proliferation and migration</td>
<td>Autoimmune demyelinating disease (123), gastric cancer (124), uterine leiomyoma and cancer (105)</td>
</tr>
<tr>
<td>229273_at</td>
<td>SALL1</td>
<td>Sal-like 1 (<em>Drosophila</em>)</td>
<td>Ureteric bud development, transcription regulation of transcription, DNA-dependent anatomical structure morphogenesis, inductive cell-cell signaling, role in cell apoptosis</td>
<td>Townes-Brocks syndrome (125)</td>
</tr>
<tr>
<td>230207_s_at</td>
<td>DOCK5</td>
<td>Guanyl-nucleotide exchange factor activity</td>
<td>Osteosarcoma, oncogenesis (126)</td>
<td></td>
</tr>
<tr>
<td>232164_s_at</td>
<td>EPPK1</td>
<td>Structural molecule activity, protein binding</td>
<td>Breast metaplastic carcinoma (127)</td>
<td></td>
</tr>
<tr>
<td>235371_at</td>
<td>GLT8D4</td>
<td>Carbohydrate biosynthetic process</td>
<td>No data available</td>
<td></td>
</tr>
<tr>
<td>241938_at</td>
<td>QKI</td>
<td>Vascularogenesis, mRNA processing, regulation of translation, transport, multicellular organismal development, spermatogenesis, axon ensheathment, RNA splicing, cell differentiation, myelination, muscle cell differentiation, long-chain fatty acid biosynthetic process, mRNA transport, role in cell apoptosis</td>
<td>Schizophrenia (128), glioma (129)</td>
<td></td>
</tr>
</tbody>
</table>

**Bold, classification between normal and tumor types.**
Table II. Genes contributing to the SVM classifier for survival classification.

<table>
<thead>
<tr>
<th>Probe set ID</th>
<th>Gene symbol</th>
<th>Gene title</th>
<th>Biological process term</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>213032_at</td>
<td>NFIB</td>
<td>Nuclear factor I/B</td>
<td>DNA replication and transcription, regulation of transcription, DNA-dependent, role in cell formation</td>
<td>Pleiomorphic adenomas (130), adipocytic tumor (131)</td>
</tr>
<tr>
<td>225423_x_at</td>
<td>LOC100129015</td>
<td>Hypothetical protein LOC100129015</td>
<td></td>
<td>Parkinson's disease (132), sarcoma formation by H-Ras oncogene-transformed human fibroblasts (133)</td>
</tr>
<tr>
<td>1554168_a_a</td>
<td>SH3KBPI</td>
<td>SH3-domain kinase binding protein 1</td>
<td>Endocytosis, apoptosis, cell-cell signaling, regulation of apoptosis, role in cell apoptosis and cell migration</td>
<td>Breast cancer (134)</td>
</tr>
<tr>
<td>204987_at</td>
<td>ITIH2</td>
<td>Inter-α (globulin) inhibitor H2</td>
<td>Hyaluronan metabolic process, serine-type endopeptidase inhibitor activity, endopeptidase inhibitor activity, role in cell binding and phagocytosis</td>
<td>Prostate cancer (135), cancer (136), Wagner syndrome (137),</td>
</tr>
<tr>
<td>204619_s_at</td>
<td>VCAN</td>
<td>Versican</td>
<td>Cell motility, cell adhesion, heart development, response to wounding, axon regeneration, multicellular organismal development, cell recognition, role in cell adhesion, apoptosis and recognition</td>
<td>Breast cancer (134)</td>
</tr>
<tr>
<td>230869_at</td>
<td>FAM155A</td>
<td>Family with sequence similarity 155, member A</td>
<td>Gene encodes protein integral to membrane</td>
<td>No data available</td>
</tr>
<tr>
<td>202748_at</td>
<td>GBP2</td>
<td>Guanylate binding protein 2, interferon-inducible</td>
<td>Immune response, role in cell contact growth inhibition</td>
<td>Colorectal carcinoma (138)</td>
</tr>
<tr>
<td>202016_at</td>
<td>MEST</td>
<td>Mesoderm specific transcript homolog (mouse)</td>
<td>Response to retinoic acid, mesoderm development, role in cell growth</td>
<td>Silver-Russel Syndrome (139), leiomyoma and endometriosis (140)</td>
</tr>
<tr>
<td>204324_s_at</td>
<td>GOLIM4</td>
<td>Golgi integral membrane protein 4</td>
<td>Transport</td>
<td>Breast cancer (141)</td>
</tr>
<tr>
<td>205363_at</td>
<td>BBOX1</td>
<td>Butyrobetaine (γ), 2-oxoglutarate dioxygenase (γ-butyrobetaine hydroxylase) 1</td>
<td>Carnitine biosynthetic process oxidation reduction</td>
<td>Refsun disease (142)</td>
</tr>
<tr>
<td>236417_at</td>
<td>MFN1</td>
<td>Mitofusin 1</td>
<td>Multicellular organismal development, mitochondrial fusion, role in cell fragmentation, length and fusion</td>
<td>Adenocarcinoma (143,144)</td>
</tr>
<tr>
<td>222474_s_at</td>
<td>TOMM22</td>
<td>Translocase of outer mitochondrial membrane 22 homolog (yeast)</td>
<td>Protein targeting to mitochondrion, transport protein transport protein import into mitochondrion intracellular protein transport across a membrane</td>
<td>Colorectal cancer (145), kidney delayed graft function (DGF), (146), globozoosperma (147)</td>
</tr>
<tr>
<td>226237_at</td>
<td>CSNK2A2</td>
<td>Casein kinase 2, α prime polypeptide</td>
<td>Protein amino acid phosphorylation, regulation of cell cycle, signal transduction, spermatid development, Wnt receptor signaling pathway</td>
<td>Colorectal cancer (145), kidney delayed graft function (DGF), (146), globozoosperma (147)</td>
</tr>
</tbody>
</table>
achieving the goals of this study. In the current study, all biopsies were collected from patients undergoing surgery. Given that these *in vivo* tests are considered to be ‘additional’ and not ‘standard’ tests for patients in pain prior to their operation, it would have been ethically ill-advised to persuade parents to agree to additional tests unless such studies were medically indicated. However, if such data are available in the future, they will be provided for inclusion in our analysis, by the assisting oncologist in the clinical data collection.

We believe that *in vivo* 2D MRS will be useful for typing inoperable tumors in the absence of biopsies or gene expression data for two reasons. Firstly, *ex vivo* MRS can be used to classify brain tumors with high sensitivity, specificity, and accuracy using only 16 metabolites (22), which is the highest number of metabolites detected with *in vivo* 2D approaches. Secondly, there is agreement between the *ex vivo* and *in vivo* MRS data obtained. Accordingly, it will be critical that we train our algorithm (which is able to handle missing data) with both MRS and gene expression data as well as with other available data to increase its discriminatory capability.

We found that certain genes were useful to subtype brain tumors (Table I) and certain other genes were useful for survival classification (Table II). These genes have been reported for other cancers or diseases (see references within Tables) but are novel to brain tumors. The ability of two genes only to discriminate between two types of brain tumors (either high grade or low grade) and metastasis was excellent (Fig. 4). It is interesting that gene 1552797_at is relevant to a stem cell marker for malignant brain tumors, cd133 (53). This suggests that our work using adult brain tumor biopsies has demonstrated that with appropriate quality control, we are able to produce meaningful data and introduce a novel classification scheme that complements and substantiates the current hypothesis of cancer stem cells (53) as a means of determining brain tumor classification and treatment.

Furthermore, our work validates and extends previous work on classification strategy analysis algorithms for both *in vivo* and *in vitro* spectra from MRS (34–40). One of the principal difficulties in such analyses is the large number of metabolites that may contribute to the spectra, each with relative intensities that can greatly vary, even in samples of the same type (34). Nonetheless, even early studies reported that the spectra of body fluids obtained with MRS are systematically different between tumor patients and healthy individuals. In many cases, successful differentiation using both linear and nonlinear methods can be made based on single resonance peaks or ratios of resonance ranges (35). More recent work on brain tumors has shown that classification according to histological type and grade is possible using similar approaches, particularly linear discriminant analysis (LDA) after feature extraction with independent components analysis (ICA) in a Bayesian framework (41) or correlation analysis and stepwise LDA (36) or using belief networks (42) or using Support Vector Machines (SVMs) (22).

Also, our approach in fusing genomics and MRS to improve the typing and prognosis of human brain tumors agrees with the notion that fusion of different sources of information can improve system performance and facilitate detection, recognition, identification, tracking, change detection, and decision-making in defense, robotics, and
medicine (58, 59). Some studies have previously described that classifiers have attempted to combine data from different sources (60-63). We believe that an efficient fusion scheme using complementary information can improve confidence.

Finally, results garnered from this study may lead to the development of new clinical tools to better assess operable cancers via tissue fingerprinting and to facilitate the distinction in vivo MRS to inoperable cancers using metabolic biomarkers to monitor anticaner therapies, in order to improve patient survival and quality of life. More importantly, they will further elucidate the biology of brain malignancy subtypes in brain tumor patients, an important step for novel drug development. Thus, these results greatly increase the overall potential for success of future studies that combine clinical MRI, MRS and MR imaging of gene expression in vivo to produce improved combined images, which could then be used to readily discriminate between metastasis and high-grade gliomas, a distinction not made adequately at present (60).

Acknowledgements

The study was supported in part by discretionary funds to A. Aria Tzika. The authors thank the Departments of Surgery and Radiology at Massachusetts General Hospital, as well as Neurosurgery and Pathology at Brigham and Women's Hospital Boston for supporting this study. We thank John Passanese, first year medical student at Harvard Medical School for assistance with references cited in Tables. We also thank Drs Ann Power Smith and Kathryn Edmondson of Write Science Right for editorial assistance.

References


