OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Multi-study Integration of Brain Cancer Transcriptomes Reveals Organ-Level Molecular Signatures

Jaeyun Sung[1,2¤], Pan-Jun Kim[3,4], Shuyi Ma[1,2], Cory C. Funk[1], Andrew T. Magis[1,5], Yuliang Wang[1,2], Leroy Hood[1], Donald Geman[6,7], Nathan D. Price[1]*

1 Institute for Systems Biology, Seattle, Washington, United States of America, **2** Department of Chemical and Biomolecular Engineering, University of Illinois, Urbana, Illinois, United States of America, **3** Asia Pacific Center for Theoretical Physics, Pohang, Gyeongbuk, Republic of Korea, **4** Department of Physics, POSTECH, Pohang, Gyeongbuk, Republic of Korea, **5** Center for Biophysics and Computational Biology, University of Illinois, Urbana, Illinois, United States of America, **6** Institute for Computational Medicine, Johns Hopkins University, Baltimore, Maryland, United States of America, **7** Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, Maryland, United States of America

Abstract

We utilized abundant transcriptomic data for the primary classes of brain cancers to study the feasibility of separating all of these diseases simultaneously based on molecular data alone. These signatures were based on a new method reported herein – Identification of Structured Signatures and Classifiers (ISSAC) – that resulted in a brain cancer marker panel of 44 unique genes. Many of these genes have established relevance to the brain cancers examined herein, with others having known roles in cancer biology. Analyses on large-scale data from multiple sources must deal with significant challenges associated with heterogeneity between different published studies, for it was observed that the variation among individual studies often had a larger effect on the transcriptome than did phenotype differences, as is typical. For this reason, we restricted ourselves to studying only cases where we had at least two independent studies performed for each phenotype, and also reprocessed all the raw data from the studies using a unified pre-processing pipeline. We found that learning signatures across multiple datasets greatly enhanced reproducibility and accuracy in predictive performance on truly independent validation sets, even when keeping the size of the training set the same. This was most likely due to the meta-signature encompassing more of the heterogeneity across different sources and conditions, while amplifying signal from the repeated global characteristics of the phenotype. When molecular signatures of brain cancers were constructed from all currently available microarray data, 90% phenotype prediction accuracy, or the accuracy of identifying a particular brain cancer from the background of all phenotypes, was found. Looking forward, we discuss our approach in the context of the eventual development of organ-specific molecular signatures from peripheral fluids such as the blood.

Citation: Sung J, Kim P-J, Ma S, Funk CC, Magis AT, et al. (2013) Multi-study Integration of Brain Cancer Transcriptomes Reveals Organ-Level Molecular Signatures. PLoS Comput Biol 9(7): e1003148. doi:10.1371/journal.pcbi.1003148

Editor: Isidore Rigoutsos, Thomas Jefferson University, United States of America

Received November 29, 2012; **Accepted** June 5, 2013; **Published** July 25, 2013

Funding: This work was supported by the National Institutes of Health/National Cancer Institute Howard Temin Pathway to Independence Award in Cancer Research (NDP), a Young Investigator Grant from the Roy J. Carver Charitable Trust (NDP), the Camille Dreyfus Teacher-Scholar Awards Program (NDP), Basic Science Research Program through the National Research Foundation of Korea (NRF-2012R1A1A2008925) (PJK), a National Science Foundation Graduate Research Fellowship (SM), the National Institutes of Health/National Center for Research Resources Grant UL1 RR 025005 (DG), and the Grand Duchy of Luxembourg-Institute for Systems Biology Program (LH, NDP). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: I have read the journal's policy and have the following conflict: The authors disclose an intent to file a patent surrounding the marker panel of the brain cancer classifiers.

* E-mail: nprice@systemsbiology.org

¤ Current address: Asia Pacific Center for Theoretical Physics, Pohang, Gyeongbuk, Republic of Korea.

Introduction

One important goal in systems medicine is to develop molecular diagnostics that can accurately and comprehensively report health and disease states of an organ system [1,2]. The discovery of organ-level molecular signatures [3] from global biomolecule expression measurements would mark a significant advance toward this goal. In this regard, genome-wide transcriptomic data are readily available, making this a promising source for molecular signatures as well as a good means to study the robustness of signatures across different studies. During the past decade, transcriptomics analyses on clinical patient samples have been widely used to uncover cancer-associated genes [4] and to discover biomarkers for diagnosis, prognosis prediction, or optimal therapy selection [5–7]. Recently, RNAs measured in blood have also been used as serum-based molecular fingerprints of neurological disease [8].

While many molecular signature studies have focused on identifying differences between case (e.g., cancer) and control (e.g., normal), a more clinically relevant and challenging task is the multi-category classification problem. This task pertains especially to identifying signatures for molecular screening and monitoring purposes. Such signatures need to detect and stratify various pathological conditions simultaneously; they must therefore be highly specific for a particular disease as well as tissue of origin. The successful identification of more reliable and efficient molecular signatures will also be critical for the blood-based, organ-specific diagnostics envisioned for the future [9].

Author Summary

From a multi-study, integrated transcriptomic dataset, we identified a marker panel for differentiating major human brain cancers at the gene-expression level. The ISSAC molecular signatures for brain cancers, composed of 44 unique genes, are based on comparing expression levels of pairs of genes, and phenotype prediction follows a diagnostic hierarchy. We found that sufficient dataset integration across multiple studies greatly enhanced diagnostic performance on truly independent validation sets, whereas signatures learned from only one dataset typically led to high error rate. Molecular signatures of brain cancers, when obtained using all currently available gene-expression data, achieved 90% phenotype prediction accuracy. Thus, our integrative approach holds significant promise for developing organ-level, comprehensive, molecular signatures of disease.

Data-driven, hierarchical approaches to multi-category classification have been investigated extensively in machine learning [10,11]. The basic idea of these methods is first to construct a classification framework in the form of a hierarchy, so that multi-category classifications can be reformulated into a series of binary decision sets (i.e., discriminating one class or group of classes from a second class or group of classes). The next step is to identify binary classifiers for all decision points (i.e., nodes and/or edges) of the hierarchy. This principle can be applied directly towards molecular disease classification, wherein all diseases can be organized into a global hierarchy of disease sets, where the diseases in each set share common expression patterns. The sets of binary classifiers can further be aggregated into a classifier marker-panel, which can direct diagnosis of an unlabeled patient sample down the hierarchical structure towards a particular label. Therefore, the cumulative expression patterns constitute ''hierarchically-structured'' molecular signatures.

A significant drawback to the use of molecular signatures derived from high-throughput—particularly transcriptomic—data is limited reproducibility and performance accuracy, which is often observed across independent studies of what are considered the same disease phenotype. This drawback holds true for both binary and multi-category classification problems. The lack of robustness, even for promising results, can be attributed to molecular heterogeneity within tumors or other diseased tissue-samples [12,13], complex disease subtypes, various patient demographics, and/or other biologically relevant factors. Another major issue is *batch effects*, which arise from differences or inconsistencies in experimental protocols, data quality, data-processing techniques, and laboratory conditions and personnel [14].

A promising method to address some of these limitations in robustness is to accumulate and combine data from many independent studies into large meta-analyses [15,16]. This integrated strategy naturally expands sample sizes across diverse sources and conditions and can therefore provide more reliable disease signatures as phenotype-associated signals become stronger relative to noise from batch effects and other sources of variance.

In this study, we developed a computational approach called Identification of Structured Signatures And Classifiers (ISSAC) to identify molecular signatures that simultaneously distinguish major cancers of the human brain. From an integrated dataset of publicly available gene expression data, ISSAC provides a global diagnostic hierarchy and corresponding structured brain cancer signatures composed of sets of gene-pair classifiers. The signal in the transcriptomics data was sufficient to develop accurate, comprehensive signatures, as long as the training set was sampled from the same population as the validation set (i.e., cross validation). In contrast, training on one dataset and testing against an independent set (i.e., an independent study measured from another lab) generally failed to reach the same performance due to biological and technical sources of dataset variation. To address this issue, we found that integration of datasets from multiple studies enhanced the disease signal sufficiently to mitigate batch effects and greatly improve independent validation results for brain cancers.

Results/Discussion

We compiled a multi-study, integrated dataset of brain cancer and normal transcriptomes [17–30] (Table 1, Table S1, and Table S2), on which we used our ISSAC algorithm (described below) to assemble classifiers into a node (Table 2, Table S3, and Figure 1) and a decision-tree (Table 3, Table S4, and Figure 2) marker panel. Importantly, while developing our algorithm to identify molecular signatures of brain cancer, we explored the effects of integrating data from multiple studies on classification performance, confirming that our integrated approach does indeed lead to more robust phenotype signatures.

Our marker panel consists of 39 total gene pairs and 44 unique genes (46 unique Affymetrix microarray probe IDs). Details on how the gene-pair sets were chosen as classifiers, and how they are used for phenotype prediction, can be found in the Materials and Methods section and Text S1. In addition, we discuss how the genes and gene pairs in our marker panel were found to have previously confirmed associations with brain cancer. Overall, we generated a marker panel with reasonably high multi-class brain cancer classification accuracy and straightforward biological interpretation.

An overview of Identification of Structured Signatures and Classifiers (ISSAC)

Here, we summarize the overall method of ISSAC into three main steps (Figure S1); a detailed algorithm and step-by-step guide are presented in the Materials and Methods section and Text S1, respectively. First, ISSAC constructs the framework for brain cancer diagnosis (Figure 3A and Figure S2)—a tree-structured hierarchy of all brain phenotypes including ependymoma (EPN), glioblastoma multiforme (GBM), medulloblastoma (MDL), meningioma (MNG), oligodendroglioma (OLG), pilocytic astrocytoma (PA), and normal brain, built using an agglomerative hierarchical clustering algorithm on gene expression training data. The construction of the hierarchy relies on iteratively identifying pairs of phenotype groups based on shared features in gene expression. As shown in Figure 3A, the cumulative set of different phenotypes is partitioned into smaller and more homogeneous subsets, thereby decomposing the multi-class diagnosis problem into more tractable sub-problems of class prediction.

Second, ISSAC identifies gene-pair classifiers corresponding to the nodes and edges of the diagnostic hierarchy (Figures 1 and 2 and Tables 2 and 3). Both types of classifiers are binary, i.e., attempt to distinguish between two sets of phenotypes. The objective of a node classifier is to distinguish the set of phenotypes associated with the node from *all other* phenotypes. For example, the classifiers of node 6 in Figure 1 and Table 2 can predict the class label of a particular transcriptome sample as either glioma (EPN, GBM, OLG, and PA) or non-glioma (MNG, MDL, and normal). In the case of an edge-based, decision-tree classifier, the objective is to distinguish the two sets of phenotypes associated with the two child nodes, analogous to rules of an ordinary decision tree. In the case of the two genes *PRPF40A* and *PURA* in

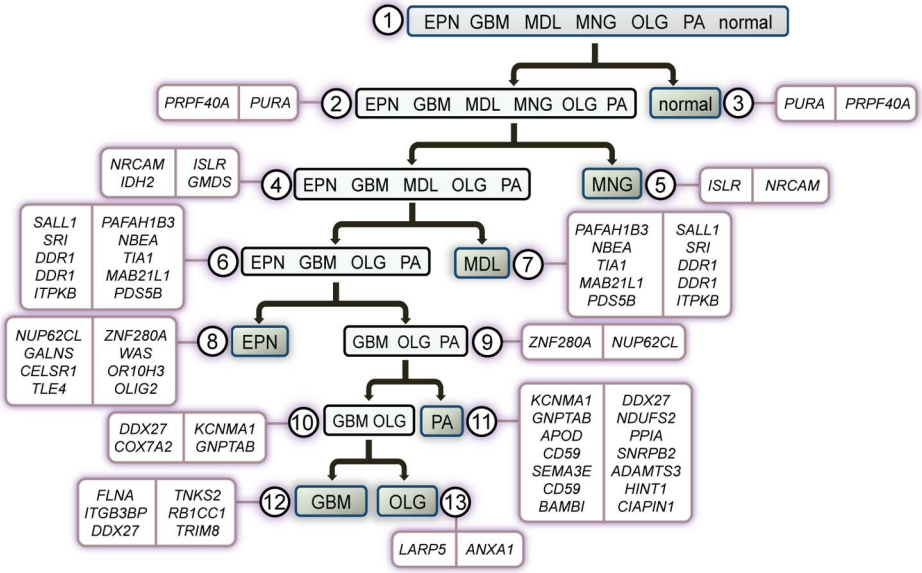

Figure 1. Gene-pair sets of the node marker-panel are shown at their corresponding twelve nodes in the brain cancer diagnostic hierarchy. Gene *i* (left) and Gene *j* (right) are the genes expressed higher and lower within each gene-pair, respectively. A transcriptome test sample is classified as the phenotype(s) of the node if the number of corresponding gene-pairs with a 'true' outcome for the statement "Gene *i* is expressed higher than Gene *j*" is greater than or equal to a threshold *k* defined for that node.
doi:10.1371/journal.pcbi.1003148.g001

Figure 2 and Table 3, this classifier determines the label of a sample as either brain cancer or normal phenotype. All classifiers are based on comparing the relative expression values (i.e., ranks) between two genes or several pairs of genes within a gene expression profile (Figures 1 and 2 and Tables 2 and 3). The chosen pairs are those that best differentiate between the phenotype sets and are based entirely on the *reversal of relative expression* (Materials and Methods), as previously reported [31].

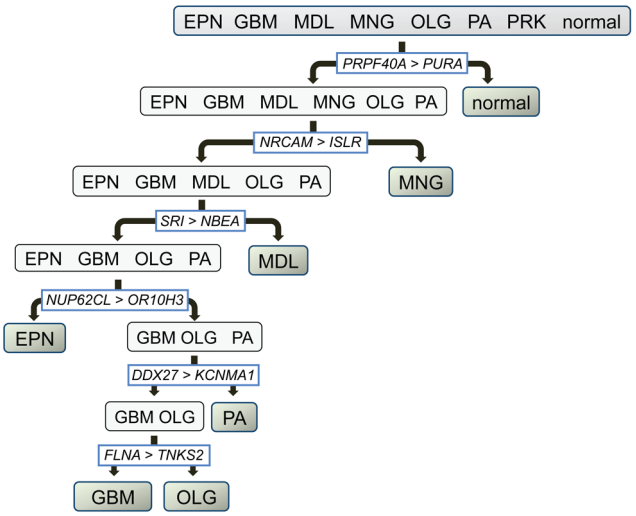

Figure 2. Gene-pairs of the decision-tree marker-panel are shown at their corresponding edges in the brain cancer diagnostic hierarchy. Gene *i* and Gene *j* are the genes expressed higher and lower within the gene-pair, respectively. For a given test sample, the direction of its classification down the diagnostic hierarchy is based on the gene-pair classifiers' true/false outcomes (left/right, respectively) for the statement "Gene *i* is expressed higher than Gene *j*".
doi:10.1371/journal.pcbi.1003148.g002

Briefly, the decision rule by Geman *et al.* is based on two genes (e.g., gene *i* and gene *j*) for distinguishing between two phenotypes (e.g., class *A* and class *B*): If the expression of gene *i* is greater than that of gene *j* for a given profile, then the phenotype is classified as class *A*; otherwise, class *B*. It has been shown that using such simple decision rules with only a small number of gene pairs can lead to highly accurate supervised classification of human cancers [32,33]. We describe the advantages of using relative expression reversals in Text S2. In addition, we provide a summary of the expression differences between classifier genes *i* and *j* in Table S5.

Overall, the collection of node classifiers represent a series of coarse-grained to fine-grained explanations of the hierarchical groupings and are used in diagnosis to screen for phenotype-specific expression patterns (described below). Thus, the hierarchy of binary predictors guides classification of an expression profile in a dynamic *coarse-to-fine* fashion: a classifier is executed if and only if all of its ancestor classifiers have been executed and have returned a positive response—i.e., predicted the phenotypes in each node. The cumulative outcome of the node classifiers for a given expression profile is the set of its candidate phenotypes, corresponding to all the leaves of the hierarchy that were reached and tested positively. This property means that it is possible to traverse multiple paths to multiple leaf nodes, and thus multiple diagnoses may be made in this step (though in practice it is usually just one). For tie-breaking purposes, the decision-tree classifiers at the edges of the diagnostic hierarchy are used to reach a unique diagnosis.

Finally, ISSAC uses the gene-pair classifiers for class prediction (Figure 3B). Given a transcriptome sample, ISSAC executes the node classifiers in a hierarchical, top-down fashion within the disease diagnostic hierarchy to identify the phenotype(s) whose class-specific signature(s) is present. As shown in Figure 3B, transcriptome samples 4–7 all have expression signatures of at least one class, i.e., a sample is classified (positive) as at least one terminal node (leaf) phenotype. In contrast, samples 1–3 do not have any class-specific signatures, i.e., samples are not positive for

Table 1. Description of all GEO microarray datasets used in this study.

Phenotype name	GEO accession #	First author (publication year)	Ref.	Sample size	Affymetrix array
Ependymoma	GSE16155	Donson (2009)	17	19	U133 plus2.0
	GSE21687	Johnson (2010)	18	83	U133 plus2.0
Glioblastoma Multiforme	GSE 4412	Freije (2004)	19	59	U133A
	GSE 4271	Phillips (2006)	20	76	U133A
	GSE 8692	Liu (2007)	21	6	U133A
	GSE 9171	Wiedemeyer (2008)	22	13	U133 plus2.0
	GSE 4290	Sun (2006)	23	77	U133 plus2.0
Medulloblastoma	GSE 10327	Kool (2008)	24	61	U133 plus2.0
	GSE 12992	Fattet (2009)	25	40	U133 plus2.0
Meningioma	GSE 4780	Scheck (2006)	-	62	U133A/U133 plus2.0
	GSE 9438	Claus (2008)	26	31	U133 plus2.0
	GSE 16581	Lee (2010)	27	68	U133 plus2.0
Oligodendroglioma	GSE 4412	Freije (2004)	19	11	U133A
	GSE 4290	Sun (2006)	23	50	U133 plus2.0
Pilocytic Astrocytoma	GSE 12907	Wong (2005)	28	21	U133A
	GSE 5675	Sharma (2007)	29	41	U133 plus2.0
Normal Brain	GSE 3526	Roth (2006)	30	146	U133 plus2.0
	GSE 7307	Roth (2007)	-	57	U133 plus2.0

Studies that have not been published are denoted as '-'.
doi:10.1371/journal.pcbi.1003148.t001

any leaf, and are labeled as "Unclassified". In case of multiple class candidates, i.e., node classifiers for multiple leaves are positive as in samples 6 and 7, the ambiguity is resolved by aggregating all the decision-tree classifiers into a classification decision-tree, thereby leading any expression signature down one unique path toward a single phenotype. Once the hierarchy and classifiers were determined, ISSAC distinguished brain cancer phenotypes with an accuracy of 90% in ten-fold cross-validation (discussed below). When the individual transcriptomic samples used in the training set were re-examined, ISSAC correctly observed all samples with an apparent (resubstitution) accuracy of 94%. This gives a sense for the relatively small degree of over-fitting compared to the cross-validation accuracy estimate.

Integrating disparate datasets identifies more robust molecular signatures across independent studies

To estimate the robustness of signature accuracy, it is best to test molecular signatures against datasets (i.e., patient samples) that are truly independent of the training set (e.g., drawn from a different patient population, clinical laboratory, etc.). To study the effects of training across multiple studies, we used glioblastoma (GBM), where we had the highest number of transcriptomic datasets for the phenotype. We trained ISSAC on each of the five transcriptomic datasets (i.e., GSE #) of GBM, coupled in each case to all the data from the other brain phenotypes. The full multi-class signatures were completely relearned (every step) with the only difference in each case being which single GBM dataset was included in the training stage. We then assessed the accuracy of correctly classifying GBM transcriptomes measured in the four held-out datasets from all other possible phenotypes. We term this evaluation method as "hold-one-lab-in validation".

The overall hold-one-lab-in validation performance, or the average of all classification accuracies in Figure 3a, was 38%. This shows that, in general, individual datasets do not consistently yield robust molecular signatures. For example, GBM signatures from GSE8692 (6 samples, ref. 21) and GSE9171 (13 samples, ref. 22) led to average accuracies of 22% and 0% for classifying independent GBM samples from other studies, respectively. These significantly low performance results are not surprising for these sets given the very small sample numbers. To an extent, relatively larger datasets could indeed yield disease signatures of higher average accuracy. However, sample size was not the sole determining factor of signature performance. For example, training on GSE4412 (59 samples, ref. 19) gave an average accuracy of 23% (Figure 4a) on the remaining GBM samples from the other studies. As a notable exception, training on GSE4271 (76 samples, ref. 20) alone resulted in the best overall average accuracy (87%) in correctly classifying samples from the four held-out GBM datasets, with individual validation set accuracies ranging from 78% to 100% (Table S6). However, when GSE4290 (77 samples, ref. 23) was used as the training set, there was over a 30% lower average GBM classification accuracy (56%) despite the nearly identical sample size with GSE4271.

We found considerable discrepancy between the minimum and maximum validation set accuracies for training sets GSE4412 (0% and 83%, respectively) and GSE4290 (17% and 92%) (Table S6). This indicates that batch effects, as well as potential biological discrepancies between populations studied at different sites, can lead to remarkable variation among transcriptomic datasets of supposedly the same phenotype. This "dataset variation" is widespread in large-scale expression studies, causing inconsistencies in molecular signature identification and performance reproducibility [34]. Large variation within and across transcriptomic datasets of GBM is perhaps not surprising, given that GBM is known to have various molecular subtypes [35]. Therefore, as mentioned above, molecular signatures from any single dataset need to be approached with caution in terms of their generalization.

Table 2. The node marker-panel is a collection of gene-pair classifiers from the nodes of the diagnostic hierarchy.

Node #[a]	Node classes[b]	Gene *i*[c]	Gene *j*[c]	*k*[d]
2	EPN GBM MDL MNG OLG PA	*PRPF40A*	*PURA*	1
3	normal	*PURA*	*PRPF40A*	1
4	EPN GBM MDL OLG PA	*NRCAM*	*ISLR*	1
		IDH2	*GMDS*	
5	MNG	*ISLR*	*NRCAM*	1
6	EPN GBM OLG PA	*SALL1*	*PAFAH1B3*	2
		SRI	*NBEA*	
		DDR1[e]	*TIA1*	
		DDR1[e]	*MAB21L1*	
		ITPKB	*PDS5B*	
7	MDL	*PAFAH1B3*	*SALL1*	4
		NBEA	*SRI*	
		TIA1	*DDR1*[e]	
		MAB21L1	*DDR1*[e]	
		PDS5B	*ITPKB*	
8	EPN	*NUP62CL*	*ZNF280A*	2
		GALNS	*WAS*	
		CELSR1	*OR10H3*	
		TLE4	*OLIG2*	
9	GBM OLG PA	*ZNF280A*	*NUP62CL*	1
10	GBM OLG	*DDX27*	*KCNMA1*	1
		COX7A2	*GNPTAB*	
11	PA	*KCNMA1*	*DDX27*	3
		GNPTAB	*NDUFS2*	
		APOD	*PPIA*	
		CD59	*SNRPB2*	
		SEMA3E	*ADAMTS3*	
		CD59	*HINT1*	
		BAMBI	*CIAPIN1*	
12	GBM	*FLNA*	*TNKS2*	1
		ITGB3BP	*RB1CC1*	
		DDX27	*TRIM8*	
13	OLG	*LARP5*	*ANXA1*	1

[a]Node # corresponds to numerical labels in the diagnostic hierarchy shown in Figure 1.
[b]Disease abbreviation (name): EPN (Ependymoma), GBM (Glioblastoma Multiforme), MDL (Medulloblastoma), MNG (Meningioma), OLG (Oligodendroglioma), PA (Pilocytic astrocytoma), and normal (Normal brain).
[c]Gene *i* and gene *j* are the genes expressed higher and lower, respectively, within each gene-pair classification decision rule. Specifically, the statement of "Gene *i* is expressed higher than Gene *j*" being true contributes to the expression profile being classified as the phenotype(s) of the node. Gene names, chromosome loci, and Affymetrix microarray platform probe IDs of the classifier genes can be found in Table S1.
[d]The minimum number of gene-pair classifiers whose decision rule outcomes for an expression profile are required to be 'true (= 1)' for the profile to be classified as the phenotype(s) of the node.
[e]Genes that share same symbol/name, but correspond to different Affymetrix probe IDs.
doi:10.1371/journal.pcbi.1003148.t002

We next analyzed how the multi-study integration approach affects performance robustness. One of each of the five datasets of GBM was sequentially withheld as the validation set, while all remaining gene expression data (including those from all other phenotypes) were used for training. The GBM signature was then evaluated on the held-out validation set. We term this strategy as "leave-one-lab-out validation". Classification accuracies using this approach ranged from 63% (GBM training set: 155 samples across four datasets; validation set: GSE4271, 76 samples) to 100% (GBM training set: 225 samples across four datasets; validation set: GSE8692, 6 samples) (Figure 4b). The average accuracy of the five leave-one-lab-out validations was 83%, which was considerably higher than that obtained from training on individual GBM datasets (38%). We conjecture that this result is due to the underlying variation in the training sets better representing the true variation in the population, both by achieving a greater sample size, as well as by having the samples come from a broader range of situations.

To evaluate how multi-study dataset integration alone affects performance robustness independent of sample size, we performed hold-one-lab-in and leave-one-lab-out validations for the studies with the largest number of samples, GSE4412, GSE4271, and GSE4290 (59, 76, and 77 samples, respectively) while training on the same number of samples for GBM. More specifically, the same steps in the analyses of Figure 4a and Figure 4b were used, while GBM signatures were learned from a training set of exactly 50 samples chosen randomly from either an individual dataset or across four combined datasets (with the fifth data set left out for validation). This process was conducted ten times for each GBM training set.

The average performances of hold-one-lab-in and leave-one-lab-out validations were 47% and 70%, respectively. Overall, the results were consistent with our two aforementioned conclusions: 1) when a molecular signature is learned from an individual dataset, its ability to accurately and precisely represent phenotype features across a broad population highly varies depending on the particular dataset used for training (Figure 4c and Table S7); and 2) combining datasets considerably increased average accuracy (Figure 4d and Table S7). Thus, dataset integration across multiple studies, even without change in sample size, can lead to significant improvements in predictive performance.

Lastly, we used the results in Figure 4c and Figure 4d to compare performances of different GBM signatures on the same validation set (Figure 4e). In all cases, signatures from combined datasets had, on average, higher classification accuracy than those from any of the individual datasets—even though the same number of samples was used in the training sets and were tested on a validation set independent of the training set. These results were then used to evaluate the precision of a GBM signature's classification accuracy by calculating its "signal-to-noise ratio (SNR)". SNR in the accuracy estimate was calculated herein as the ratio of average classification accuracy to standard deviation in the accuracy estimate across studies. We found that, for all validation set cases, GBM signatures developed on the basis of multiple datasets had SNRs greater by at least two fold than those from individual data sets. This clearly shows that learning on integrated datasets leads to molecular signatures that have higher and more consistent (i.e. less variable) predictive performance (Figure 4f), and motivated our choice in developing the brain cancer ISSAC signature to only use cases where we had at least 2 independent studies to learn across.

Overall, we have shown that when a broader range of conditions within a particular phenotype is presented during the classifier-learning stage, ISSAC can better distinguish the true disease signal from noise prior to independent validation. However, single and/or smaller training sets that were used to define the classifiers might not be representative of, or general-

Table 3. The decision-tree marker-panel shows phenotype-specific signatures in the form of binary patterns.

Gene symbols[a]		Disease binary signatures[b]						
Gene *i*	**Gene *j***	**EPN**	**GBM**	**MDL**	**MNG**	**OLG**	**PA**	**normal**
PRPF40A	*PURA*	1	1	1	1	1	1	0
NRCAM	*ISLR*	1	1	1	0	1	1	-
SRI	*NBEA*	1	1	0	-	1	1	-
NUP62CL	*OR10H3*	1	0	-	-	0	0	-
DDX27	*KCNMA1*	-	1	-	-	1	0	-
FLNA	*TNKS2*	-	1	-	-	0	-	-

[a]Affymetrix microarray platform probe IDs of the classifier genes are shown in Table S2.
[b]For each gene-pair comparison (i.e., Is Gene *i*>Gene *j* ?), 1 and 0 delineates 'true' and 'false', respectively, and '–' denotes that the outcome is not used for classification.
doi:10.1371/journal.pcbi.1003148.t003

izable to, larger populations – leading to poor validation results. Therefore, the utilization of all currently available datasets from various sources and conditions may be a promising approach to finding novel diagnostic markers, and eventually bringing the successful adaptation of genomic biomarkers into clinical practice. Also, prospective design of studies is generally best when they utilize multiple sites to avoid over-fitting to particular contexts.

It is worth mentioning that in some cases, molecular signatures from a single source can have (or at least appear to have) superior performance, as demonstrated by the molecular signatures from GSE4271. Specifically, training on a single GSE4271 data set provided higher accuracy (87%, Figure 4a) than learning on any of the four sets combined (average 83%, Figure 4b). Indeed, when such surprisingly robust single datasets are identified, they potentiate significant new insight into the underlying heterogeneities present in a patient population of a disease phenotype. Such data sets can be utilized for follow-up studies, and hence serve as a valuable resource to the scientific and medical communities. It is, however, difficult in practice to predict in advance data set robustness, which must be ensured through careful sample collection and data set preprocessing techniques. To help ensure the production of reliable omics-based data sets, we recommend the following: 1) Good experimental design, such as clearly defining clinical phenotypes of interest; 2) When collecting new experimental data, sufficient sample size must be obtained; 3) All aspects of the experimental and analytical procedures must be carefully controlled to avoid batch effects; and 4) No confounding from factors unrelated to phenotype(s) of interest must occur.

Brain cancer marker-panel achieves high classification accuracy in cross-validation

As shown by our leave-one-lab-out validations, learning signatures across multiple datasets significantly improved average classification accuracy with concomitant reduction in performance variance. In this regard, the brain cancer marker-panel obtained using all currently available microarray data simultaneously (Tables 2 and 3) should represent more robust phenotype signatures.

The classification performance of this comprehensive brain cancer marker-panel was evaluated by ten-fold cross-validation (Figure S3). Our marker-panel achieved a 90% average of phenotype-specific classification accuracies (Table 4), showing strong promise against a multi-category, multi-dataset background at the gene expression level. In addition, we observed higher classification accuracy (93%) among the expression profiles for which a unique diagnosis was obtained without subsequent disambiguation from the decision-tree (Table S8). Furthermore, the glioblastoma (GBM) classification accuracy previously seen in our leave-one-lab-out analysis (83%) is comparable to that seen in cross-validation (85%). Indeed, that these two accuracies are so close suggests that, for GBM, the effects of variability among the datasets from different institutions and time-points have been mostly overcome by integration across multiple training studies.

Four other brain cancers (ependymoma, medulloblastoma, meningioma, and pilocytic astrocytoma) have estimated accuracies of at least 91%, suggesting clear differences between them and the other phenotypes at the transcriptomic level. The anatomical region specificity of these four cancers may have contributed toward their highly accurate separation, as there are regional areas of unique gene expression patterns. Roth *et al.* analyzed gene expression of 20 anatomically distinct regions of the central nervous system [30] and clustered all anatomical sites into distinct groups, providing evidence of region-specific expression patterns. However, results from another study analyzing gene expression data from distinct brain regions suggested that clustering disparities might also be due to activity of distinct brain cell types, rather than solely on region [36,37]. Furthermore, if region specificity played a dominant role in classification, we would expect to see a high number of misdiagnoses to occur between the normal brain, which was derived from 25 different locations (Text S3), and the six cancers. Such a trend was not observed in Table 4. Therefore, our predictive results suggest a stronger contribution from underlying cell-type specific and disease-intrinsic elements than from region effects alone.

Compared to the cross-validation accuracies of other phenotypes, lower performance was observed for GBM and oligodendroglioma (OLG) (85% and 75%, respectively). This could have been mainly a consequence of the limited ability of the marker-panel to correctly differentiate these two cancers from each other. Indeed, the distinction of these two phenotypes from transcriptomics seems to be rather difficult in general, and our accuracies here are comparable to those reported previously in two-phenotype comparison studies [38,39]. Furthermore, our signatures did show an excellent degree of sensitivity (96%) and specificity (97%) for distinguishing these two well-progressed gliomas as a set from all other brain phenotypes. There exist genetic tests and methods that differentiate GBM and OLG well, such as the combined loss of chromosome arms 1p and 19q [40], and over-expression of the transcription factor protein Olig2 [41], but our goal in this particular study was to evaluate molecular discriminatory power as represented in transcriptomes across multiple brain cancers.

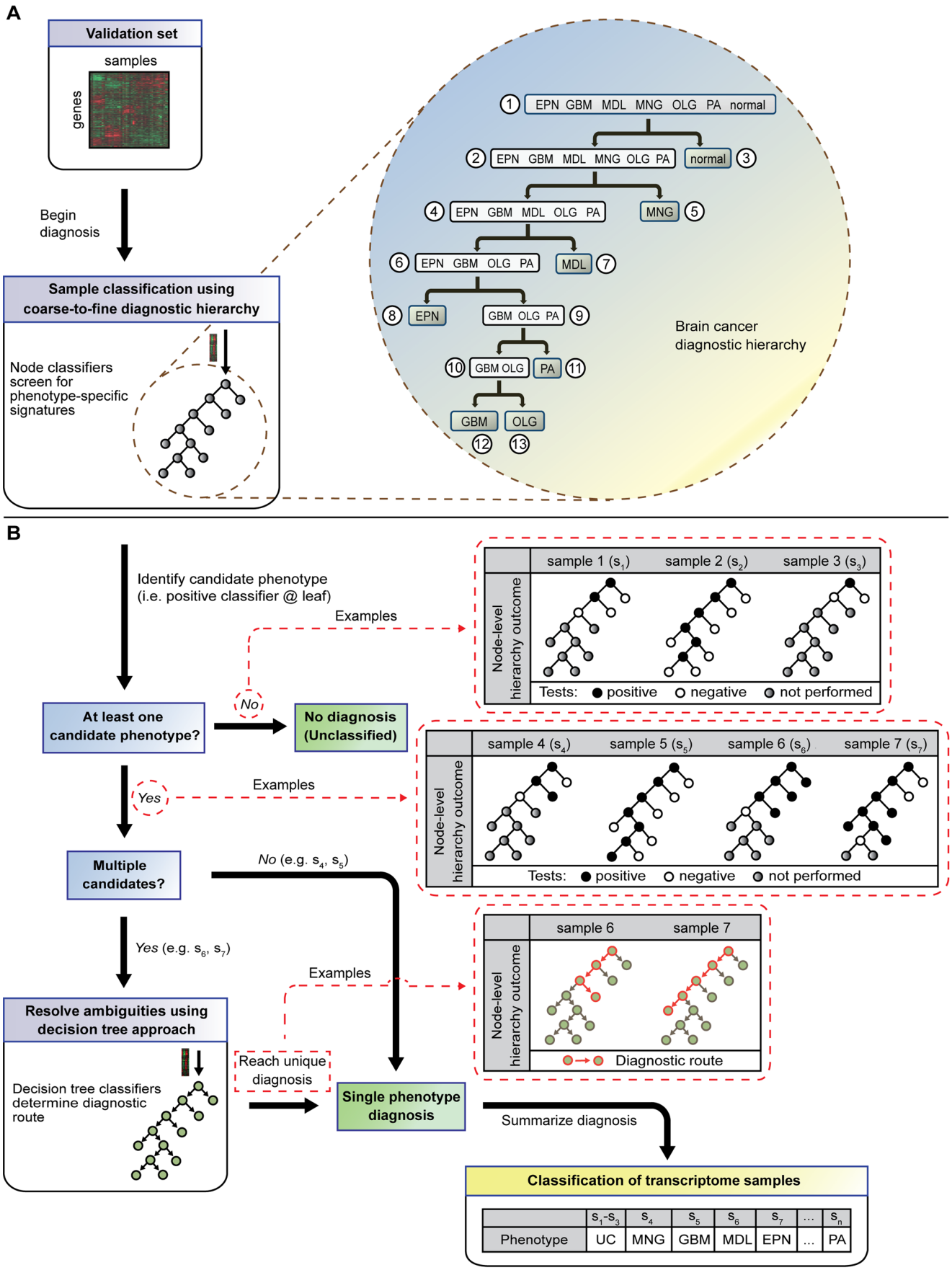
A
Validation set
samples
genes
Begin diagnosis
Sample classification using coarse-to-fine diagnostic hierarchy
Node classifiers screen for phenotype-specific signatures
1 EPN GBM MDL MNG OLG PA normal
2 EPN GBM MDL MNG OLG PA
normal 3
4 EPN GBM MDL OLG PA
MNG 5
6 EPN GBM OLG PA
MDL 7
8 EPN
GBM OLG PA 9
10 GBM OLG
PA 11
GBM
OLG
12 13
Brain cancer diagnostic hierarchy
B
Identify candidate phenotype (i.e. positive classifier @ leaf)
Examples
At least one candidate phenotype?
No
No diagnosis (Unclassified)
Yes
Examples
Multiple candidates?
No (e.g. s4, s5)
Yes (e.g. s6, s7)
Examples
Resolve ambiguities using decision tree approach
Decision tree classifiers determine diagnostic route
Reach unique diagnosis
Single phenotype diagnosis
Summarize diagnosis
Node-level hierarchy outcome
sample 1 (s1) sample 2 (s2) sample 3 (s3)
Tests: positive negative not performed
sample 4 (s4) sample 5 (s5) sample 6 (s6) sample 7 (s7)
Tests: positive negative not performed
sample 6 sample 7
Diagnostic route
Classification of transcriptome samples
Phenotype | s1-s3 | s4 | s5 | s6 | s7 | ... | sn
UC | MNG | GBM | MDL | EPN | ... | PA

Figure 3. Comprehensive classification of human brain cancer and normal brain transcriptomes using molecular signatures from ISSAC. A The coarse-to-fine classification process is represented by a hierarchically structured groupings of phenotypes. There is a node classifier for each set of phenotypes in the hierarchy, which is designed to respond positively if the sample belongs to this set of diseases and negatively otherwise. Our diagnostic hierarchy has thirteen nodes in total, and seven terminal nodes (i.e., leaves). The node classifiers are executed sequentially and adaptively on a given expression profile; a classifier test for a particular node is performed if and only if all of its ancestor tests were performed and deemed positive. The node classifiers are used to screen for phenotype-specific signatures. **B** Leaves that have positive classifier outcomes correspond to the candidate phenotypes of a given expression profile. If there is no candidate phenotype, the expression profile is labeled as 'Unclassified'. If only one candidate phenotype is identified, the profile is labeled as that phenotype of the respective leaf. If the profile is considered to consist of multiple phenotype signatures, the ambiguity is resolved using the decision-tree classifiers based on the same diagnostic hierarchy. Here, the decision-tree classifiers are executed starting from the root of the tree, directing the profile to one of the two child nodes sequentially until it completes a full path towards a leaf. The phenotype label of the final destination corresponds to the unique diagnosis.
doi:10.1371/journal.pcbi.1003148.g003

Marker-panel genes' association to cancer biology

Several genes in our marker panel are strongly associated with brain cancers, suggesting putative relationships to the underlying pathophysiology of their corresponding phenotypes. One such gene is *NRCAM* (nodes 4 and 5 of Figure 1 and Table 2), which was reported as a marker for high-risk neuroblastoma [42] and poor prognostic ependymoma [43]. *NRCAM* was also found to be over-expressed in cell lines derived from pilocytic astrocytomas and glioblastoma multiforme tumors [44]. The receptor tyrosine kinase *DDR1*, a predicted marker gene for PA when expressed higher than *TIA1* and *MAB21L1* (nodes 6 and 7), was found to be over-expressed in high-grade gliomas and to promote tumor cell invasion [45]. *FLNA* was detected in the serum of high-grade astrocytoma (grade 3 and GBM) patients [46], and *ANXA1*, a gene that encodes an anti-inflammatory phospholipid binding protein, was implicated in astrocytoma progression [47]. These reports are consistent with our identification of *FLNA* and *ANXA1* as two classifier genes expressed higher in GBM than in oligodendroglioma (nodes 12 and 13). The basic helix-loop-helix (bHLH) transcription factor *OLIG2* is innately expressed in oligodendrocytes and was recently characterized as a key antagonist of p53 function in neural stem cells and malignant gliomas [48]. In accordance with lower expression of *OLIG2* as an EPN classifier in this study (node 8), *OLIG2* expression was used as a negative marker to differentiate EPN from other gliomas [49]. *SEMA3E*, one of several classifier genes for PA (node 11), has been reported to drive invasiveness of melanoma cells in mice [50]. And finally, mutation to *IDH2* (node 4) in GBM is well known, with occurrence reported in 80% of secondary glioblastomas [51,52]. That the genes in our marker panel have previously confirmed ties to brain cancers raises the question of what is the underlying molecular framework surrounding the generation of gene-pair classifiers, which would be an interesting direction for future studies. Among the gene pairs in our marker panel, we focus on two pairs (below) in which the genes' common functional roles or relevance to cancer suggest putative relationships to corresponding pathology. Our discussions below point to potential biological relationships underlying the observed gene expression reversals, representing hypotheses that require further experimental validation.

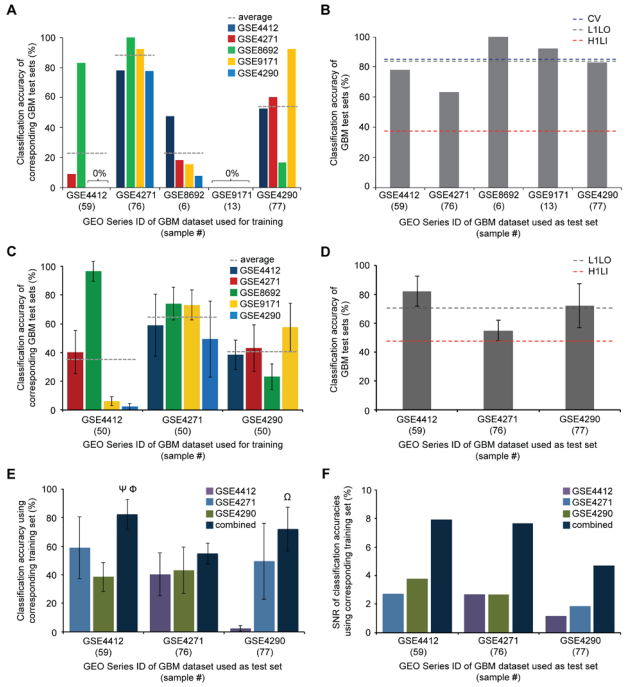

Figure 4. Molecular signatures from multi-study, integrated datasets have higher average phenotype prediction accuracy and lower performance variance than those from individual datasets. A Hold-one-lab-in validation results for each of the five glioblastoma (GBM) datasets. Gray line indicates average accuracy on the four validation sets. **B** Leave-one-lab-out validation results for each of the five GBM datasets. Blue and red line indicates average accuracy of GBM signatures from leave-one-lab-out (L1LO) validation and hold-one-lab-in (H1LI) validation, respectively. **C** H1L1 validation to test GBM signatures from GSE4412, GSE4271, and GSE4290, while keeping the number of samples in the GBM training set the same. 50 samples were randomly selected from each GBM dataset for signature learning. H1LI validation was executed ten times for each of the three GBM datasets. Error bars indicate standard deviations. **D** L1LO validation to test GBM signatures on GSE4412, GSE4271, and GSE4290 validation sets, while 50 total samples were randomly selected from the other four GBM datasets for signature learning. L1LO validation was executed ten times. **E** Data from **C** and **D** are used to show GBM signatures' accuracies on GSE4412, GSE4271, and GSE4290 validation sets when the GBM training data were from individual or combined GBM datasets. Ψ, Φ, and Ω indicate statistical significance relative to GSE4271, GSE4290, and GSE4412, respectively (Tukey's post-hoc test, $p<0.01$). **F** Signal-to-noise ratios (SNRs) from data in **E**. SNR was calculated as the ratio of average accuracy to standard deviation.
doi:10.1371/journal.pcbi.1003148.g004

One of the classifier gene pairs involved in the differentiation between meningioma and the remaining five brain cancers (EPN, GBM, MDL, OLG, PA) are two metabolic enzymes, *IDH2* and *GMDS* (node 4). *IDH2* converts isocitrate to α-ketoglutarate within the TCA cycle. This reaction produces NADPH, which not only is an essential cofactor for many metabolic reactions, but also helps to protect the cell against oxidative damage [53]. Moreover, *GMDS* aids the biosynthesis of GDP-fucose from GDP-mannose in mannose metabolism, in which NADPH is produced [54]. That the enzymatic activities of both *IDH2* and *GMDS* participate in the conversion between $NADP^+$ and NADPH is interesting, considering the well-known alteration to cellular metabolism and deregulated redox balance in cancer [55]. Possible MNG-specific mutations in *IDH2* and/or *GMDS*, or changes in the regulatory network that controls the expression of these two genes, may affect cellular redox balance and functions of other metabolic enzymes.

Table 4. Classification performance of brain cancer marker-panel in ten-fold cross-validation.

Actual phenotype	Predicted phenotype (%)[a]								Total
	EPN	GBM	MDL	MNG	OLG	PA	normal	UC[b]	
EPN	**92.2**	2.8	0.3	1.7	1.3	0.6	0.2	1.0	102
GBM	0.7	**84.8**	0.2	0.5	11.9	0.1	0.3	1.3	231
MDL	2.2	2.3	**91.1**	0.8	2.7	0.2	0.0	0.8	101
MNG	0.1	1.8	0.0	**97.5**	0.1	0.2	0.0	0.2	161
OLG	0.5	20.7	0.2	0.0	**74.6**	2.1	0.0	2.0	61
PA	1.3	2.3	0.0	0.0	1.3	**94.4**	0.0	0.8	62
normal	0.0	0.5	0.0	0.1	0.7	0.0	**98.5**	0.1	203

[a]Accuracies reflect average performance in ten-fold cross-validation conducted ten times. The main diagonal gives the average classification accuracy of each class (bold), and the off-diagonal elements show the erroneous predictions.
[b]UC (Unclassified samples). When using the node classifiers, expression profiles that did not exert a signature of any phenotype (i.e., did not percolate down to at least one positive terminal node) were rejected from classification. In this case, the Unclassified sample is treated as a misclassification.
doi:10.1371/journal.pcbi.1003148.t004

The *TLE4* and *OLIG2* gene pair is used to differentiate EPN from GBM, OLG, and PA (node 8). *TLE4*, a human homolog of the Drosophila Groucho protein, represses the Wnt and FGF developmental signaling pathways [56–58] by recruiting deacetylases to histones H3 and H4 [59]. FGF receptor signaling was reported to control neuronal and glial cell development by regulating *OLIG2* expression in zebrafish [58]. This connection between these two genes in regards to brain cell development could be reflective of the extent of cell-type differentiation (a hallmark of cancer), or lack thereof, unique to EPN compared with the other gliomas.

To develop further hypotheses of the functional relationships between the classifiers and pathophysiological traits, we looked for statistical enrichment of biological properties (e.g. biological processes, chromosome numbers) on an exhaustive list of gene pairs discriminating GBM and OLG (Text S4). Our statistical enrichment of biological processes of gene-set *i* and gene-set *j* (the union of genes in each gene-pair classifier that are expressed relatively higher and lower in GBM, respectively) showed that the genes reflect disease properties. Specifically, the genes that are in gene-set *i*, or those expressed higher in GBM compared to OLG, were the most enriched in the biological process of 'Immunity and Defense' (Figure S4); this is in concordance with clinical observations showing high degree of inflammation inside malignant tumors (such as GBM), as well as the subsequent high number of immune cells. Our additional reports on the statistical enrichment of certain chromosome numbers link our classifiers to known genomic aberrations of their respective brain cancers, providing further insight as to why certain genes might have been selected as classifiers.

Looking ahead: Molecular signatures based on putative blood borne biomolecules offer a glimpse into possible molecular diagnostics

The work reported herein has focused on identifying a structured molecular signature that can separate major brain cancers simultaneously, as well as on evaluating issues related to reproducibility in molecular signatures. However, our long-term motivation for wanting molecular signatures of an organ system is ultimately to find corresponding signatures in the blood, where they can be assayed non-invasively. Blood bathes virtually all organs, which secrete proteins and nucleic acids. Subsets of these secreted biomolecules can potentially constitute disease signatures for molecular diagnostics, as measurement technologies mature. Moreover, the blood is easily accessible in contrast to biopsies of diseased organs for obtaining transcript or protein profiles. In this regard, the brain represents an organ system where a critical need exists to develop non-invasive techniques to monitor its health state through secreted proteins.

Previously, organ-specific proteins have been detected in blood; when these proteins changed in concentration or chemical structure, the tissue origin of this change was identified [60]. For blood-based, organ-specific diagnostics, molecular signatures need to detect and stratify various possible cancers and other pathological conditions simultaneously. In the context of this current study, an intriguing question is if training ISSAC on shed or secreted blood borne biomolecule measurements identifies molecular signatures that allow us to distinguish health from disease; and if diseased, which one and how far has it progressed? Thus, the approach laid out herein for transcriptomics is a foundation for identifying similar signatures from blood proteins as these measurements become more abundant.

As proof of concept and to provide candidates for targeted proteomics analysis, we performed the above transcriptomic analysis of finding brain cancer signatures using only the genes that are annotated to encode extracellular proteins (Materials and Methods). We trained ISSAC on a total of 767 genes that matched this criterion, which led to a new brain cancer marker-panel composed of 41 gene-pair classifiers from 71 unique features (Figure S5). When looking at the case of GBM gene-pair classifiers, i.e. 59 node-based genes involved in the detection of either GBM or phenotype groups that include GBM, 11 were previously identified as potential GBM-specific serum markers (detected either from GBM cell-line secretome experiments or in human plasma): *APOD* [61], *CALU* [62], *CD163* [63,64], *CHI3L1* [65–67], *CSF1* [68,69], *EGFR* [68,70,71], *IGFBP2* [62,72–75], *NID1* [76], *PDGFC* [77,78], *PSG9* [72], and *PTN* [79]. We provide the functional roles of these genes in Table S9. None of these previous studies performed relative abundance comparisons or measured expression ratios, so we are unable to answer at this time whether the particular relative expression reversal patterns would be observed in serum. We were not able to find any direct available evidence associating the remaining GBM classifier genes to potential serum-based markers. Nevertheless, we were encouraged that ISSAC was able to verify some previously identified potential GBM markers, which provides support for its use towards a blood-

based test since there is currently no clinically approved GBM-specific, serum-based biomarker.

Our marker-panel, composed entirely of genes encoding extracellular products, obtained an average classification accuracy of 87% in 10-fold cross-validation (Table S10), which compares favorably to the average accuracy we previously achieved using all the genes in the microarray (90%). This suggests that strong signal may possibly persist for phenotype distinction even when using only secreted biomolecules from diseased organs. If indeed there are enough biomolecules secreted into the blood at concentrations that can be accurately and consistently detected by e.g., targeted mass spectrometry, then there is the very exciting possibility that organ-specific pathologies, such as those described above, can be detected from the blood. This would truly make blood a powerful window into health and disease.

Materials and Methods

Multi-study dataset of human brain cancer transcriptomes

All transcriptomic data used in our analysis are publicly available at the NCBI Gene Expression Omnibus (GEO). We integrated 921 microarray samples of six brain cancers (ependymoma, glioblastoma multiforme, medulloblastoma, meningioma, oligodendroglioma, pilocytic astrocytoma) and normal brain across 16 independent studies into a transcriptome multi-study dataset. Importantly, we obtained the raw data (.CEL files) from each of these studies and preprocessed them uniformly using identical techniques to greatly reduce extraneous sources of technical artifacts (discussed below). All data manipulation and numerical calculations were performed using MATLAB (MathWorks).

To ensure data quality and to help control for systemic bias and batch effects), we used the following strict criteria and reasoning for brain phenotype selection: 1) Expression profiles must have been conducted on either the Affymetrix Human Genome U133A or U133 Plus 2.0 microarray platform. This allowed maximum microarray sample collection without considerable reduction in number of overlapping classifier features (i.e., microarray probe-sets). 2) Transcriptomic datasets (i.e., GSE #) for each phenotype must have been collected from at least two independent sources to help mitigate batch effects. 3) All datasets must have consisted of no fewer than 5 microarray samples. 4) All datasets must have originated from primary brain tumor or tissue biopsies. Expression profiles from cell-lines or laser micro-dissections were not used in our study to better ensure sample consistency. 5) Raw microarray intensity data (.CEL files) must have been available on GEO for consensus preprocessing (described below). 6) Sample preparation protocols must have been fully disclosed on GEO. 7) All microarray samples in a dataset of a given phenotype were used in order to take into consideration all sources of heterogeneity. That is, *no* samples were excluded because their gene expression profiles were abnormal for their associated phenotypes. We are aware that this may allow mislabeled samples, e.g. samples that were originally misclassified by the histopathologist upon class labeling (Text S5), to be used in the classifier-learning stage, and thereby limit the biological "purity" of a phenotype in the training set. This can pose a serious challenge in interpreting misclassified samples that actually seem to be a much better match (or even perfect match) to a different phenotype, leading to questions of whether a misclassification is due to ISSAC's limitation in distinguishing phenotypes, or whether a re-evaluation of the original tumor biopsy is required. Despite these concerns, we concluded this to be the most stringent test. After an exhaustive search on GEO, we identified 921 microarray samples from 16 studies that met the above criteria (as of January 2011). Information on all datasets (e.g., publication sources, Affymetrix platforms, GEO dataset IDs, and microarray sample IDs), studies, and GEO microarray sample IDs used in our study is available in Table 1, Table S1, and Table S2, respectively.

Raw microarray intensity data (.CEL files) were obtained online from GEO and preprocessed uniformly. More specifically, common probe-sets were found across all transcriptome samples, and consensus preprocessing was performed on all the raw microarray image data to build a consensus dataset. This step removes one major non-biological source of variance between different studies. These preprocessed samples were used to build a multi-study integrated dataset of human brain cancer and normal brain transcriptomes. Finally, stringent probe-set filtering was used to remove spurious classifier features. Our consensus preprocessing and probe-set filtering methods are explained in further detail below. Our integrated and uniformly pre-processed dataset is available on our group's webpage (http://price.systemsbiology.net/downloads) as a community resource for those who wish to conduct their own analyses.

Consensus preprocessing using GCRMA

All gene expression data used in our study were measurements conducted on either the Affymetrix Human Genome U133A or U133Plus2.0 oligonucleotide microarrays. The expression level of a target gene on these two platforms is measured by first quantifying the total intensity of fluorescently labeled RNA fragments (from patient specimens) that bind to a probe set, or the set of complementary 25-mer oligonucleotide probe sequences. The intensities of all probe sets (raw measurements in the form of .CEL files) are then adjusted for background variability and normalized across all samples to obtain the target genes' final expression values.

Raw .CEL data files were downloaded directly from GEO. Probe set information used in this study were based on the latest Affymetrix annotations. Raw intensity measurements of all microarray samples considered in our study were preprocessed collectively (consensus preprocessing) using the MATLAB implementation of the microarray preprocessing GCRMA [80]. Only the probe sets that map to known genes and exist on both Affymetrix platforms (same oligonucleotide sequences) were considered for preprocessing. The use of individual Affymetrix probe sets as classifiers (and not the mean or median of their expression values as demonstrated in other microarray-based studies) imposes limitations in the classifiers' multi-platform compliance, as discussed in Text S6 and Text S7.

Probe set filtering using MAS5 calls

Probe sets of Affymetrix microarrays have "perfect match" probes that are exactly complementary to the target gene's mRNA sequence. They also have "mismatch" probes that contain a mismatched nucleotide halfway along the probe sequence, and are used to estimate the degree of non-specific binding. To ensure that a probe set is reliably detected, the measurements of the "perfect match" probes must be significantly greater than those of the "mismatch" probes. This is usually assessed based on statistical measures. The MAS5 preprocessing software makes expression quality calls based on the nonparametric Wilcoxon signed-rank test. The "absent" call is made when the p-value is greater than 0.06, representing no significant difference between the measurements of the "perfect match" and those of the 'mismatch' probes [81]. We eliminated probes that were determined to be "absent" in all samples of the consensus dataset. After this probe set filtering

step, 19,656 probe sets (corresponding to target genes) within each microarray sample were kept for further analysis.

All GCRMA preprocessing and MAS5 probe set filtering procedures were conducted separately for training and test set samples, i.e., inside each cross-validation or hold-out loop, in order to avoid possible cross-talk between the two datasets. Genes that were excluded based on the MAS5 "absent" calls on the training data were also removed from the corresponding test data.

Description of ISSAC algorithm

A tree-structured framework for brain cancer diagnostics. Let $\mathcal{L}$ denote the set of class labels, in our case the seven brain phenotypes: six cancers and normal. Given an expression profile x, the objective is to determine its true phenotype $\Upsilon \in \mathcal{L}$.

The main assumption is that there are natural groupings $L \subset \mathcal{L}$ among the phenotypes. Thus, testing for these groupings can more efficiently utilize the available training data, leading to more accurate classification than testing for each phenotype individually. Based on these attributes, the natural structure to represent $\mathcal{L}$ is then a diagnostic hierarchy in the form of a binary hierarchical decision tree T. Each node $t \in \mathcal{T}$ is associated with a set of phenotypes $L_t \subset \mathcal{L}$. The root of T contains all the phenotypes and each leaf (terminal node) of T delineates a single phenotype. Overall, this representation is nested, in the sense that the set of phenotypes at every non-terminal node is the disjoint union of the phenotypes of the two child nodes. This tree is built from the training data by agglomerative hierarchical clustering derived from features of the profiles, as discussed below.

Node classifiers are assembled according to the diagnostic hierarchy. There is a binary classifier f_t for every node $t \in \mathcal{T}$ except for the root. The classifier f_t is a function of the expression profile x. Put simply, f_t is a collective "test" for phenotypes in L_t versus all other phenotypes. More formally, the classifier returns two possible outcomes: $f_t(\mathrm{x}) = 1$ (i.e., positive) signals that we accept the hypothesis that $\Upsilon \in L_t$, and $f_t(\mathrm{x}) = 0$ (i.e., negative) signals that we reject this hypothesis and conclude that $\Upsilon \notin L_t$. In particular, f_t is *not* a test for L_t versus the phenotypes in the sibling of t, as would be the case with a standard decision tree. Rather, f_t looks for traits within a given profile x which characterize all phenotypes in L_t *simultaneously*, such that a positive result signifies that the classifier assumes the true class of x belongs to L_t.

Classifier learning begins at the two child nodes of the root, and the classifiers are learned from two types of training data. The positive training data for learning the classifier f_t for node t are all the expression profiles of the phenotypes in L_t. The negative training data are all the profiles of the phenotypes that are not in L_t.

Being binary, each classifier has two performance metrics: the *sensitivity* of f_t is the probability that $f_t(\mathrm{x}) = 1$ given x is from the positive training data, and the *specificity* of f_t is the probability that $f_t(\mathrm{x}) = 0$ given x is from the negative training data. Due to the coarse-to-fine, hierarchical manner in which the classifiers are processed, we required the *sensitivity* of f_t to be as close to unity as possible. This can be accomplished at the expense of specificity by adjusting a threshold, as discussed below. The reason for imposing a high sensitivity on each classifier is that if a test profile is rejected from belonging to L_t by the classifier when in fact it does belong to L_t, it cannot be recovered. However, the reduced specificity is only local to each node, and the overall specificity increases with testing at subsequent nodes.

A coarse-to-fine screening yields candidate phenotypes. The strategy for processing any given profile x with the diagnostic hierarchy is breadth-first, coarse-to-fine. Starting from the two child nodes of the root, classifiers are executed sequentially and adaptively, with f_t performed if and only if all its ancestor tests are performed and are positive. More specifically, f_t is performed if and only if $f_s = 1$ for every node $s \in T$ between t and the root. As soon as $f_t = 0$ for a non-terminal node t, none of the descendant classifiers in the sub-tree rooted at t are performed. This is because a negative response of f_t means that the phenotype is unlikely to belong to L_t and the set of phenotypes associated with descendant of t, which are necessarily subsets of L_t. This facilitates pruning whole subsets of phenotypes at once.

The complete coarse-to-fine screening process for x results in a *set* of detected phenotypes. We denote this set by $L(\mathrm{x}) \subset \mathcal{L}$. These are the phenotypes corresponding to a complete chain of positive results for all f_t from root to leaf. Equivalently, $L(\mathrm{x})$ is the set of phenotypes that are not ruled out by any test performed. During the diagnostic process, a profile may traverse only one path all the way to the terminal node. In this case, $L(\mathrm{x})$ consists of a single phenotype d, and the diagnostic process terminates with $\Upsilon = d$ as the predicted phenotype. However, a profile may also traverse multiple branches to the terminal nodes of T, in which case $L(\mathrm{x})$ consists of multiple candidate phenotypes (see the discussion on resolving ambiguities below). Moreover, a profile may reach no terminal nodes, in which case $L(\mathrm{x})$ is empty. When no terminal node is reached, the profile is determined to be outside of L, and labeled as 'Unclassified'.

Resolving ambiguities using a decision-tree approach. When $L(\mathrm{x})$ consists of multiple phenotypes, it becomes necessary to refine the diagnosis. The ambiguity is resolved by another tree-structured process – an ordinary decision tree based on the same diagnostic hierarchy. For every pair of sibling nodes, e.g., $L_t = \{EPN\}$ and $L_s = \{GBM, OLG, PA\}$, we learn a classifier $g_{t,s}$ which tests $\Upsilon \in L_t$ versus $\Upsilon \in L_s$, just as in an ordinary decision-tree (the process of classifier identification is elaborated below). Starting from the root of the tree, execution of the decision-tree classifiers directs a profile to one of two sibling nodes sequentially until it reaches a terminal node. Unlike the process of traversing the hierarchy of node classifiers, a sample that enters the decision tree is directed to one and only one leaf node, and hence uniquely labeled.

Classifier design and learning. Every node classifier is based on expression level comparisons between two genes. Let G be the set of all genes for which we have microarray expression data, and denote the set of all distinct *pairs* of genes by $\mathcal{P}$. For each gene-pair $(g_i, g_j) \in \mathcal{P}$, consider the Boolean feature $Z_{ij}(\mathrm{x}) \in \{0,1\}$ of an expression profile $\mathrm{x} = \{x_g,\ g \in G\}$. $Z_{ij}(\mathrm{x})$ assumes the value 1 if gene g_i is expressed higher than gene g_j (i.e., $x_{g_i} > x_{g_j}$) in x, and the value 0 otherwise (i.e., $x_{g_i} \le x_{g_j}$). These are the features that have been used in previous work on relative expression reversals [31]. Each node classifier f is constructed from a small set of gene-pairs $P \subset \mathcal{P}$, the binary outcomes of Z_{ij} for all $(i,j) \in P$, and a constant threshold k. More specifically, $f(\mathrm{x}) = 1$ if $\sum_{(g_i,g_j) \in P} Z_{ij} \ge k$, and $f(\mathrm{x}) = 0$ otherwise. The threshold k takes values between 1 and $|P|$.

There is a classifier of this nature for every node $t \in T$, except for the root. The set of gene-pairs $P = P_t$ and threshold $k = k_t$ depend on the node t. Hence, for each t, the classifier $f_t = 1$ if at least k_t of the gene-pair comparisons in P_t are positive $(Z_{ij} = 1)$; otherwise, $f_t = 0$. The comparisons are chosen such that, for each pair (g_i, g_j) in P_t, we expect to see gene g_i expressed more than gene g_j under the assumption that the phenotype of x belongs to L_t, whereas if the phenotype of x does not belong to L_t, we expect to see the reverse. For every node t, every pair of all gene-pair combinations

is "scored" by the difference between the probability of the event that $Z_{ij} = 1$ given $Y \in L_t$ and the probability given $Y \notin L_t$. These probabilities are estimated from the training data, and the subset of pairs with the highest scores are chosen.

Since each positive (resp., negative) comparison is viewed as evidence for $Y \in L_t$ (resp., $Y \notin L_t$), we can then favor sensitivity over specificity by varying the threshold k_t. That is, by choosing a relatively small value for k_t relative to the number of comparisons in P_t, we can make it highly likely that the classifier responds positively when in fact the sample belongs to the set L_t. We show the sets of gene-pairs P_t for each of the twelve nodes in our diagnostic hierarchy in Table 2 and an illustrative example in Figure 1. Finally, the decision tree classifiers $g_{t,s}$ are all based on comparisons of *single* gene-pairs at all edges of the diagnostic hierarchy.

While multiple gene pairs were used at each decision point in the node-based tree, only a single gene pair was used at each decision point in the decision-tree. This is due to the difference in the motivation of building the two trees; the node-based tree was constructed to maximize sensitivity and minimize false-positives with as many pairs as necessary, while the decision-tree was designed to resolve multiple diagnoses (i.e. ties) which could be done with only one pair. We show the pair of genes for each of the six decision-tree classifiers in Table 3 and Figure 2. MATLAB implementations of the ISSAC algorithm and a step-by-step tutorial are available to download at http://price.systemsbiology.net/downloads.

Selecting genes that encode extracellular products

Using Gene Ontology (GO) annotations, we have identified a list of 767 genes (mapped on 1,085 total probes) in every transcriptome sample that encode for possible blood-borne proteins. Specifically, we selected only the genes whose products are annotated to be in either the 'Extracellular Space' or the 'Extracellular Region' cellular locations. We use this gene set as a starting point for targeted blood diagnostics. All computational steps and analyses in regards to molecular signature discovery are identical to those discussed above.

Supporting Information

Figure S1 The overall method of ISSAC can be summarized into three main steps. **A** ISSAC constructs the framework for brain cancer diagnosis – a tree-structured hierarchy of all brain cancer phenotypes built using an agglomerative hierarchical clustering algorithm on gene expression training data. **B** Training on gene-expression data from all brain phenotypes, ISSAC identifies disjoint, gene-pair classifiers at all nodes (excluding the root) and edges of the diagnostic hierarchy, and accumulates them into their respective marker panels. The chosen pairs are the ones that best differentiate between the phenotype sets, and are based entirely on the *reversal of relative expression*. **C** ISSAC uses the gene-pair classifiers for class prediction. Briefly, given a gene expression profile, ISSAC executes the node classifiers in a hierarchical, top-down fashion within the disease diagnostic hierarchy to identify the phenotype(s) whose class-specific signature(s) is present. In case of multiple class candidates (i.e. node classifiers for multiple leaves are positive), the ambiguity is resolved by aggregating all the decision-tree classifiers into a classification decision-tree, thereby leading any expression signature down one unique path toward a single phenotype.
(TIFF)

Figure S2 Brain phenotypes are grouped into a global diagnostic hierarchy, which allows an intuitive representation of the classification process. The diagnostic hierarchy is built using a data-driven, iterative approach, and is free of manual, ad-hoc construction. In each iteration, two classes, or two groups of classes, with the lowest TSP score (Materials and Methods and Text S1) among all pair-wise comparisons, come together to form a node. This approach optimizes overall classification by placing the more challenging decisions further away from the base of the tree (i.e. root), thereby ensuring only the minimum misclassifications percolate down the tree. The final form of the brain phenotype diagnostic hierarchy represents a hierarchical structure of nested partitions, where the multi-class problem is decomposed into smaller and smaller groups using a sequence of diagnostic decision rules.
(TIFF)

Figure S3 Performance evaluation using ten-fold cross-validation. **A** Ten-fold cross-validation is conducted ten times to obtain the average accuracy. In every iteration of cross-validation, the order of samples within a particular class are randomly permuted before training/test set allocations. **B** Our marker panel achieved a 90.4% average of phenotype-specific classification accuracies, showing strong promise against a multi-category, multi-dataset background at the gene expression level.
(TIFF)

Figure S4 Statistical enrichment analysis on PANTHER database biological processes and chromosome numbers of the top 500, 1,000, and 1,500 gene-pair classifiers for GBM vs. OLG. **A** 'Immunity and Defense' was the most enriched biological process for 'gene-set *i*', reflecting the chronic inflammatory conditions inside the GBM tumor. **B** 'Neuronal Activities' was the most enriched biological process for 'gene-set *j*', reflecting decrease in neuronal behavior and possibly other brain cell activity inside the GBM tumor. The genes in 'gene-set *i*' and 'gene-set *j*' were the most enriched in **C** Chromosome 1 and **D** Chromosome 10, respectively, reflecting the major chromosome aberrations of the two brain cancers. Ψ delineates the most enriched category. [a]Biological process abbreviation (name): AAM (Amino acid metabolism), APT (Apoptosis), CM (Carbohydrate metabolism), CA (Cell adhesion), CC (Cell cycle), CPD (Cell proliferation and differentiation), CSM (Cell structure and motility), DP (Developmental processes), HMS (Homeostatis), IMD (Immunity and defense), IPRT (Intracellular protein transport), MCN (Muscle contraction), NA (Neuronal activities), NAM (Nucleic acid metabolism), PMM (Protein metabolism and modification), SM (Sulfur metabolism), ST (Signal transduction), and TR (Transport).
(TIFF)

Figure S5 Gene-pair classifiers based on only the genes that encode extracellular products. Gene pairs are shown at their corresponding nodes in the brain disease diagnostic hierarchy. The corresponding node-based marker panel consists of 41 classifier pairs and 71 unique classifier features.
(TIFF)

Table S1 Phenotype specimen descriptions and main results for all GEO accessions used in this study.
(PDF)

Table S2 GEO microarray sample IDs used in this study.
(PDF)

Table S3 Node marker panel for brain cancer and normal transcriptome classification. Node #: Corresponds to numerical labels shown in the brain phenotype diagnostic hierarchy (Figure 1). Brain phenotype abbreviation (name): EPN (Ependy-

moma), GBM (Glioblastoma multiforme), MDL (Medulloblastoma), MNG (Meningioma), normal (Normal brain), OLG (Oligodendroglioma), and PA (Pilocytic astrocytoma). Gene *i*/Gene *j*: the gene expressed higher and lower in the gene-pair, respectively, within each corresponding phenotype. Gene name/Chromosome locus: according to Entrez Gene. Affymetrix Probe ID: For both Affymetrix Human Genome U133A and U133Plus2.0 Arrays. *k*: The minimum number of gene-pair classifiers whose decision rule outcomes for a test sample are required to be 'true (= 1)' for the sample to be classified as the phenotype(s) of the corresponding node.
(PDF)

Table S4 Decision-tree marker panel for brain cancer and normal transcriptome classification. For each classifier decision rule (i.e. Is Gene *i*>Gene *j* ?), 1 and 0 delineates 'true' and 'false', respectively, and '–' denotes that the outcome is not used for classification. The vertical binary pattern under each class label corresponds to a phenotype-specific molecular signature.
(PDF)

Table S5 Summary of expression differences between genes-pair classifiers. Node #: Corresponds to numerical labels shown in the brain phenotype diagnostic hierarchy (Figure 1). Brain phenotype abbreviation (name): EPN (Ependymoma), GBM (Glioblastoma multiforme), MDL (Medulloblastoma), MNG (Meningioma), normal (Normal brain), OLG (Oligodendroglioma), and PA (Pilocytic astrocytoma). Sample number: Number of total samples in classes of respective Node #. Gene *i*/Gene *j*: the gene expressed higher and lower in the gene-pair, respectively, within each corresponding phenotype. Gene name/Chromosome locus: according to Entrez Gene. Affymetrix Probe ID: For both Affymetrix Human Genome U133A and U133Plus2.0 Arrays. *k*: The minimum number of gene-pair classifiers whose decision rule outcomes for a test sample are required to be 'true (= 1)' for the sample to be classified as the phenotype(s) of the corresponding node. Ranked expression differences of each gene pair (i.e. Rank_gene_*i* – Rank_gene_*j*) were calculated for each sample, and Mean, St. dev., Max., Min., and Median were found across all samples within classes of respective Node #.
(PDF)

Table S6 Hold-one-lab-in validation accuracies of glioblastoma signatures.
(PDF)

Table S7 Hold-one-lab-in (H1LI) and leave-one-lab-out (L1LO) validation accuracies of glioblastoma signatures when training data were constrained to 50 total samples. H1LI and L1LO validations were performed ten times for each category of training data. In each validation trial, 50 samples were randomly selected from the single microarray dataset (for H1L1) or from the multi-study, combined dataset (for L1LO).
(PDF)

Table S8 Ten-fold cross-validation accuracies when only the node marker panel was required to reach unique diagnoses. **Sample size:** Average proportion of total samples that reached unique diagnoses via node marker panel. **Accuracy:** Reflects average performance in ten-fold cross-validation conducted ten times.
(PDF)

Table S9 Functional roles of 11 previously identified GBM serum markers that are present in our extracellular-product encoding marker-panel. [a]http://www.ncbi.nlm.nih.gov/gene.
(PDF)

Table S10 Ten-fold cross-validation accuracies of gene-pair classifiers composed of genes that encode extracellular products. [a]Accuracies reflect average performance in ten-fold cross-validation conducted ten times. The main diagonal gives the average classification accuracy of each class (bold), and the off-diagonal elements show the erroneous predictions. [b]UC (Unclassified samples). When using the node classifiers, expression profiles that did not exert a signature of any phenotype (i.e., did not percolate down to at least one positive terminal node) were rejected from classification. In this case, the Unclassified sample is treated as a misclassification.
(PDF)

Text S1 Step-by-step description of how ISSAC works.
(PDF)

Text S2 Advantages of using relative expression reversals to build classifiers.
(PDF)

Text S3 Twenty-five anatomical regions of the human brain from which normal transcriptome samples were obtained.
(PDF)

Text S4 Global statistical enrichment analysis of gene-pair classifiers.
(PDF)

Text S5 Whether ISSAC can play a role in identifying misdiagnoses.
(PDF)

Text S6 Reasoning for selecting only Affymetrix microarray platforms and for not using probe-specific offsets.
(PDF)

Text S7 Candidates of brain cancer molecular signatures.
(PDF)

Acknowledgments

We would like to thank James A. Eddy and Matthew C. Gonnerman for helpful discussions.

Author Contributions

Conceived and designed the experiments: JS DG NDP. Performed the experiments: JS. Analyzed the data: JS PJK CCF LH DG NDP. Wrote the paper: JS DG NDP. Performed all pre-processing of the raw microarray data files: ATM. Improved the MATLAB code: SM YW.

References

1. Hood L, Friend SH (2011) Predictive, personalized, preventive, participatory (P4) cancer medicine. Nat Rev Clin Oncol 8: 184–187.
2. Tian Q, Price ND, Hood L (2011) Systems Cancer Medicine: Towards Realization of Predictive, Preventive, Personalized, and Participatory (P4) Medicine. J Intern Med 271: 111–21.
3. Sung J, Wang Y, Chandrasekaran S, Witten DM, Price ND (2012) Molecular signatures from omics data: from chaos to consensus. Biotechnol J 7: 946–957.
4. Gu CC, Rao DC, Stormo G, Hicks C, Province MA (2002) Role of gene expression microarray analysis in finding complex disease genes. Genetic Epidemiology 23: 37–56.
5. Friedman DR, Weinberg JB, Barry WT, Goodman BK, Volkheimer AD, et al. (2009) A Genomic Approach to Improve Prognosis and Predict Therapeutic Response in Chronic Lymphocytic Leukemia. Clinical Cancer Research 15: 6947–6955.

6. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, et al. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine 7: 673–679.
7. Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, et al. (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell 1: 133–143.
8. Scherzer CR, Eklund AC, Morse LJ, Liao ZX, Locascio JJ, et al. (2007) Molecular markers of early Parkinson's disease based on gene expression in blood. Proceedings of the National Academy of Sciences of the United States of America 104: 955–960.
9. Hood L, Heath JR, Phelps ME, Lin BY (2004) Systems biology and new technologies enable predictive and preventative medicine. Science 306: 640–643.
10. Blanchard G, Geman D (2005) Hierarchical testing designs for pattern recognition. Annals of Statistics 33: 1155–1202.
11. Amit Y, Geman D, Fan XD (2004) A coarse-to-fine strategy for multiclass shape detection. Ieee Transactions on Pattern Analysis and Machine Intelligence 26: 1606–1621.
12. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, et al. (2011) Tumour evolution inferred by single-cell sequencing. Nature 472: 90–94.
13. Park SY, Gonen M, Kim HJ, Michor F, Polyak K (2010) Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype. The Journal of clinical investigation 120: 636–644.
14. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, et al. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics 11: 733–739.
15. Miller JA, Horvath S, Geschwind DH (2010) Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. Proceedings of the National Academy of Sciences of the United States of America 107: 12698–12703.
16. Dudley JT, Tibshirani R, Deshpande T, Butte AJ (2009) Disease signatures are robust across tissues and experiments. Molecular Systems Biology 5: 307.
17. Donson AM, Birks DK, Barton VN, Wei Q, Kleinschmidt-Demasters BK, et al. (2009) Immune gene and cell enrichment is associated with a good prognosis in ependymoma. J Immunol 183: 7428–7440.
18. Johnson RA, Wright KD, Poppleton H, Mohankumar KM, Finkelstein D, et al. (2010) Cross-species genomics matches driver mutations and cell compartments to model ependymoma. Nature 466: 632–636.
19. Freije WA, Castro-Vargas FE, Fang Z, Horvath S, Cloughesy T, et al. (2004) Gene expression profiling of gliomas strongly predicts survival. Cancer Res 64: 6503–6510.
20. Phillips HS, Kharbanda S, Chen R, Forrest WF, Soriano RH, et al. (2006) Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. Cancer Cell 9: 157–173.
21. Liu T, Papagiannakopoulos T, Puskar K, Qi S, Santiago F, et al. (2007) Detection of a microRNA signal in an in vivo expression set of mRNAs. PLoS One 2: e804.
22. Wiedemeyer R, Brennan C, Heffernan TP, Xiao Y, Mahoney J, et al. (2008) Feedback circuit among INK4 tumor suppressors constrains human glioblastoma development. Cancer Cell 13: 355–364.
23. Sun L, Hui AM, Su Q, Vortmeyer A, Kotliarov Y, et al. (2006) Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. Cancer Cell 9: 287–300.
24. Kool M, Koster J, Bunt J, Hasselt NE, Lakeman A, et al. (2008) Integrated genomics identifies five medulloblastoma subtypes with distinct genetic profiles, pathway signatures and clinicopathological features. PLoS One 3: e3088.
25. Fattet S, Haberler C, Legoix P, Varlet P, Lellouch-Tubiana A, et al. (2009) Beta-catenin status in paediatric medulloblastomas: correlation of immunohistochemical expression with mutational status, genetic profiles, and clinical characteristics. J Pathol 218: 86–94.
26. Claus EB, Park PJ, Carroll R, Chan J, Black PM (2008) Specific genes expressed in association with progesterone receptors in meningioma. Cancer Res 68: 314–322.
27. Lee Y, Liu J, Patel S, Cloughesy T, Lai A, et al. (2010) Genomic landscape of meningiomas. Brain Pathol 20: 751–762.
28. Wong KK, Chang YM, Tsang YT, Perlaky L, Su J, et al. (2005) Expression analysis of juvenile pilocytic astrocytomas by oligonucleotide microarray reveals two potential subgroups. Cancer Res 65: 76–84.
29. Sharma MK, Mansur DB, Reifenberger G, Perry A, Leonard JR, et al. (2007) Distinct genetic signatures among pilocytic astrocytomas relate to their brain region origin. Cancer Res 67: 890–900.
30. Roth RB, Hevezi P, Lee J, Willhite D, Lechner SM, et al. (2006) Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. Neurogenetics 7: 67–80.
31. Geman D (2004) Classifying gene expression profiles from pairwise mRNA comparisons. Statistical Applications in Genetics and Molecular Biology 3: Article 19.
32. Tan AC, Naiman DQ, Xu L, Winslow RL, Geman D (2005) Simple decision rules for classifying human cancers from gene expression profiles. Bioinformatics 21: 3896–3904.
33. Price ND, Trent J, El-Naggar AK, Cogdell D, Taylor E, et al. (2007) Highly accurate two-gene classifier for differentiating gastrointestinal stromal tumors and leiomyosarcomas. Proceedings of the National Academy of Sciences of the United States of America 104: 3414–3419.
34. Allison DB, Cui XQ, Page CP, Sabripour M (2006) Microarray data analysis: from disarray to consolidation and consensus (vol 7, pg 55, 2006). Nature Reviews Genetics 7: 406–406.
35. Mischel PS, Shai R, Shi T, Horvath S, Lu KV, et al. (2003) Identification of molecular subtypes of glioblastoma by gene expression profiling. Oncogene 22: 2361–2373.
36. Khaitovich P, Muetzel B, She XW, Lachmann M, Hellmann I, et al. (2004) Regional patterns of gene expression in human and chimpanzee brains. Genome Research 14: 1462–1473.
37. Oldham MC, Konopka G, Iwamoto K, Langfelder P, Kato T, et al. (2008) Functional organization of the transcriptome in human brain. Nature Neuroscience 11: 1271–1282.
38. Kim S, Dougherty ER, Shmulevich I, Hess KR, Hamilton SR, et al. (2002) Identification of combination gene sets for glioma classification. Molecular Cancer Therapeutics 1: 1229–1236.
39. Nutt CL, Mani DR, Betensky RA, Tamayo P, Cairncross JG, et al. (2003) Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. Cancer Research 63: 1602–1607.
40. Burger PC, Minn AY, Smith JS, Borell TJ, Jedlicka AE, et al. (2001) Losses of chromosomal arms 1p and 19q in the diagnosis of oligodendroglioma. A study of paraffin-embedded sections. Modern Pathology 14: 842–853.
41. Ligon KL, Alberta JA, Kho AT, Weiss J, Kwaan MR, et al. (2004) The oligodendroglial lineage marker OLIG2 is universally expressed in diffuse gliomas. Journal of Neuropathology and Experimental Neurology 63: 499–509.
42. Liu X, Mazanek P, Dam V, Wang Q, Zhao H, et al. (2008) Deregulated Wnt/beta-catenin program in high-risk neuroblastomas without MYCN amplification. Oncogene 27: 1478–1488.
43. Lukashova-v Zangen I, Kneitz S, Monoranu CM, Rutkowski S, Hinkes B, et al. (2007) Ependymoma gene expression profiles associated with histological subtype, proliferation, and patient survival. Acta Neuropathol 113: 325–337.
44. Sehgal A, Boynton AL, Young RF, Vermeulen SS, Yonemura KS, et al. (1998) Cell adhesion molecule Nr-CAM is over-expressed in human brain tumors. Int J Cancer 76: 451–458.
45. Weiner HL, Huang H, Zagzag D, Boyce H, Lichtenbaum R, et al. (2000) Consistent and selective expression of the discoidin domain receptor-1 tyrosine kinase in human brain tumors. Neurosurgery 47: 1400–1409.
46. Alper O, Stetler-Stevenson WG, Harris LN, Leitner WW, Ozdemirli M, et al. (2009) Novel anti-filamin-A antibody detects a secreted variant of filamin-A in plasma from patients with breast carcinoma and high-grade astrocytoma. Cancer Sci 100: 1748–1756.
47. Schittenhelm J, Trautmann K, Tabatabai G, Hermann C, Meyermann R, et al. (2009) Comparative analysis of annexin-1 in neuroepithelial tumors shows altered expression with the grade of malignancy but is not associated with survival. Mod Pathol 22: 1600–1611.
48. Mehta S, Huillard E, Kesari S, Maire CL, Golebiowski D, et al. The Central Nervous System-Restricted Transcription Factor Olig2 Opposes p53 Responses to Genotoxic Damage in Neural Progenitors and Malignant Glioma. Cancer Cell 19: 359–371.
49. Ishizawa K, Komori T, Shimada S, Hirose T (2008) Olig2 and CD99 are useful negative markers for the diagnosis of brain tumors. Clin Neuropathol 27: 118–128.
50. Casazza A, Finisguerra V, Capparuccia L, Camperi A, Swiercz JM, et al. Sema3E-Plexin D1 signaling drives human cancer cell invasiveness and metastatic spreading in mice. J Clin Invest 120: 2684–2698.
51. Yan H, Parsons DW, Jin G, McLendon R, Rasheed BA, et al. (2009) IDH1 and IDH2 mutations in gliomas. N Engl J Med 360: 765–773.
52. Gross S, Cairns RA, Minden MD, Driggers EM, Bittinger MA, et al. Cancer-associated metabolite 2-hydroxyglutarate accumulates in acute myelogenous leukemia with isocitrate dehydrogenase 1 and 2 mutations. J Exp Med 207: 339–344.
53. Reitman ZJ, Yan H (2010) Isocitrate dehydrogenase 1 and 2 mutations in cancer: alterations at a crossroads of cellular metabolism. Journal of the National Cancer Institute 102: 932–941.
54. Becker DJ, Lowe JB (2003) Fucose: biosynthesis and biological function in mammals. Glycobiology 13: 41R–53R.
55. Hanahan D, Weinberg RA (2011) Hallmarks of Cancer: The Next Generation. Cell 144: 646–674.
56. Ahn SM, Byun K, Kim D, Lee K, Yoo JS, et al. (2008) Olig2-induced neural stem cell differentiation involves downregulation of Wnt signaling and induction of Dickkopf-1 expression. PLoS One 3: e3917.
57. Burks PJ, Isaacs HV, Pownall ME (2009) FGF signalling modulates transcriptional repression by Xenopus groucho-related-4. Biol Cell 101: 301–308.
58. Esain V, Postlethwait JH, Charnay P, Ghislain J (2010) FGF-receptor signalling controls neural cell diversity in the zebrafish hindbrain by regulating olig2 and sox9. Development 137: 33–42.
59. Winkler CJ, Ponce A, Courey AJ Groucho-mediated repression may result from a histone deacetylase-dependent increase in nucleosome density. PLoS One 5: e10166.
60. Lausted C, Hu Z, Hood L (2008) Quantitative serum proteomics from surface plasmon resonance imaging. Mol Cell Proteomics 7: 2464–2474.

61. Gautam P, Nair SC, Gupta MK, Sharma R, Polisetty RV, et al. (2012) Proteins with altered levels in plasma from glioblastoma patients as revealed by iTRAQ-based quantitative proteomic analysis. Plos One 7: e46153.
62. Somasundaram K, Nijaguna MB, Kumar DM (2009) Serum proteomics of glioma: methods and applications. Expert Rev Mol Diagn 9: 695–707.
63. Persson A, Englund E (2012) Phagocytic properties in tumor astrocytes. Neuropathology 32: 252–260.
64. Komohara Y, Ohnishi K, Kuratsu J, Takeya M (2008) Possible involvement of the M2 anti-inflammatory macrophage phenotype in growth of human gliomas. J Pathol 216: 15–24.
65. Gollapalli K, Ray S, Srivastava R, Renu D, Singh P, et al. (2012) Investigation of serum proteome alterations in human glioblastoma multiforme. Proteomics 12: 2378–2390.
66. Formolo CA, Williams R, Gordish-Dressman H, MacDonald TJ, Lee NH, et al. (2011) Secretome signature of invasive glioblastoma multiforme. J Proteome Res 10: 3149–3159.
67. Polisetty RV, Gupta MK, Nair SC, Ramamoorthy K, Tiwary S, et al. (2011) Glioblastoma cell secretome: analysis of three glioblastoma cell lines reveal 148 non-redundant proteins. J Proteomics 74: 1918–1925.
68. Coniglio SJ, Eugenin E, Dobrenis K, Stanley ER, West BL, et al. (2012) Microglial stimulation of glioblastoma invasion involves epidermal growth factor receptor (EGFR) and colony stimulating factor 1 receptor (CSF-1R) signaling. Mol Med 18: 519–527.
69. Ryder M, Gild M, Hohl TM, Pamer E, Knauf J, et al. (2013) Genetic and pharmacological targeting of CSF-1/CSF-1R inhibits tumor-associated macrophages and impairs BRAF-induced thyroid cancer progression. Plos One 8: e54302.
70. Quaranta M, Divella R, Daniele A, Di Tardo S, Venneri MT, et al. (2007) Epidermal growth factor receptor serum levels and prognostic value in malignant gliomas. Tumori 93: 275–280.
71. Heimberger AB, Suki D, Yang D, Shi W, Aldape K (2005) The natural history of EGFR and EGFRvIII in glioblastoma patients. J Transl Med 3: 38.
72. Sreekanthreddy P, Srinivasan H, Kumar DM, Nijaguna MB, Sridevi S, et al. (2010) Identification of potential serum biomarkers of glioblastoma: serum osteopontin levels correlate with poor prognosis. Cancer Epidemiol Biomarkers Prev 19: 1409–1422.
73. Lin Y, Jiang T, Zhou K, Xu L, Chen B, et al. (2009) Plasma IGFBP-2 levels predict clinical outcomes of patients with high-grade gliomas. Neuro Oncol 11: 468–476.
74. Li Y, Jiang T, Zhang J, Zhang B, Yang W, et al. (2012) Elevated serum antibodies against insulin-like growth factor-binding protein-2 allow detecting early-stage cancers: evidences from glioma and colorectal carcinoma studies. Ann Oncol 23: 2415–2422.
75. Fukushima T, Kataoka H (2007) Roles of insulin-like growth factor binding protein-2 (IGFBP-2) in glioblastoma. Anticancer Res 27: 3685–3692.
76. Gupta MK, Polisetty RV, Ramamoorthy K, Tiwary S, Kaur N, et al. (2013) Secretome analysis of Glioblastoma cell line - HNGC-2. Mol Biosyst 9: 1390–400.
77. di Tomaso E, London N, Fuja D, Logie J, Tyrrell JA, et al. (2009) PDGF-C induces maturation of blood vessels in a model of glioblastoma and attenuates the response to anti-VEGF treatment. Plos One 4: e5123.
78. di Tomaso E, Snuderl M, Kamoun WS, Duda DG, Auluck PK, et al. (2011) Glioblastoma recurrence after cediranib therapy in patients: lack of ''rebound'' revascularization as mode of escape. Cancer Res 71: 19–28.
79. Poimenidi E, Hatziapostolou M, Papadimitriou E (2009) Serum stimulates Pleiotrophin gene expression in an AP-1-dependent manner in human endothelial and glioblastoma cells. Anticancer Res 29: 349–354.
80. Wu Z IR, Gentleman R, Martinez-Murillo F, Spencer F (2004) A model-based background adjustment for oligonucleotide expression arrays. Journal of the American Statistical Association 99: 909–917.
81. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4: 249–264.