Cancer Research

Correlation of Somatic Mutation and Expression Identifies Genes Important in Human Glioblastoma Progression and Survival

David L. Masica and Rachel Karchin

Abstract

Cooperative dysregulation of gene sequence and expression may contribute to cancer formation and progression. The Cancer Genome Atlas (TCGA) Network recently catalogued gene sequence and expression data for a collection of glioblastoma multiforme (GBM) tumors. We developed an automated, model-free method to rapidly and exhaustively examine the correlation among somatic mutation and gene expression and interrogated 149 GBM tumor samples from the TCGA. The method identified 41 genes whose mutation status is highly correlated with drastic changes in the expression (z-score $\pm$ 2.0), across tumor samples, of other genes. Some of the 41 genes have been previously implicated in GBM pathogenesis (e.g., *NF1*, *TP53*, *RB1*, and *IDH1*) and others, while implicated in cancer, had not previously been highlighted in studies using TCGA data (e.g., *SYNE1*, *KLF6*, *FGFR4*, and *EPHB4*). The method also predicted that known oncogenes and tumor suppressors participate in GBM via drastic over- and underexpression, respectively. In addition, the method identified a known synthetic lethal interaction between TP53 and PLK1, other potential synthetic lethal interactions with TP53, and correlations between IDH1 mutation status and the overexpression of known GBM survival genes. *Cancer Res; 71(13); 4550–61. ©2011 AACR.*

Introduction

Cancer is a complex disease arising from the concerted effect of multiple (epi)genetic changes that yield pathway dysregulation via alterations in gene copy number, DNA methylation, gene expression, and molecular function (1–3). Specific combinations of these alterations can vary, even in histologically similar cancers. Until recently, the limited scalability of genetic experiments forbade complete characterization of these complexities and variances; now, large-scale cancer genomics experiments can catalogue alterations with up to full-exome coverage across tens to hundreds of samples (1, 2, 4). Bioinformatics techniques can interrogate this data and identify alterations that cooperatively drive cancer (5–10), even with patient specificity (6).

Alterations that affect gene expression levels [e.g., copy number alteration (CNA) and DNA methylation] in cancer genomes have been used to identify driver genes and molecular subtypes of a particular cancer (11, 12). Such alterations have identified oncogenes activated via increased expression [as can occur with epidermal growth factor receptor (EGFR), for instance] or tumor suppressors deactivated via decreased expression (as can occur with RB1, for instance). Verhaak and colleagues showed that 4 clinically relevant glioblastomas [glioblastoma multiforme (GBM)] subtypes could be defined using a subset of The Cancer Genome Atlas (TCGA) GBM expression data (11). In that study, the authors also grouped mutation and CNA with the expression-defined GBM subtypes. Increased expression can further be used to identify cancer-specific essential genes, oncogene addiction, and synthetic lethality (13, 14).

Understanding subtype and patient-specific combinatorial patterns of (epi)genetic alterations in tumors has promise to inform therapeutic regimens. First, expression patterns common to a subtype may be informative with respect to the drugs most suitable for a group of patients. For example, the neural GBM subtype has a high rate of EGFR and ERBB2 overexpression, but patients with neural GBM that are not EGFR and/or ERBB2 positive may not benefit from receptor tyrosine kinase inhibitors. Second, alterations in off-target genes can modulate the efficacy of targeted therapies (i.e., drug resistance). For instance, EGFR-positive tumors respond to gefitinib, but amplification of the MET proto-oncogene can cause resistance (15). Tumors overexpressing ERBB2 respond to trastuzumab, but phosphoinositide-3-kinase (PI3K) mutation can cause trastuzumab resistance (15). Finally, cancer-specific essential

Authors' Affiliation: Department of Biomedical Engineering and Institute for Computational Medicine, The Johns Hopkins University, Baltimore, Maryland

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

Corresponding Author: Rachel Karchin, Johns Hopkins University, 217A CSEB, 3400 N. Charlest St., Baltimore, MD 21218. Phone: 410-516-5578; Fax: 410-516-5294; E-mail: karchin@jhu.edu

doi: 10.1158/0008-5472.CAN-11-0180

genes, oncogene addiction, and synthetic lethality can be druggable vulnerabilities in tumors (13, 14). Notably, while current methods for synthetic lethal screening can identify such vulnerabilities, some studies suggest that considering isolated pairwise interactions limits generalizability. For example, 3 groups screened unique KRAS-driven cancers for synthetic lethal interactions and recovered 3 unique lists of genes synthetic lethal with *KRAS* mutation (16); this suggests that the identified synthetic lethal interactions were a subset of larger, more complex networks (i.e., context specificity).

Many current bioinformatics approaches for assessing complex patterns of (epi)genetic aberrations in cancer rely on pre-existing knowledge of gene annotations, gene sets, protein–protein interactions, and curated pathways. Gene set enrichment analysis is a widely used method for interpreting differential gene expression levels, based on previously described functions and pathway memberships. Vaske and colleagues have used CNA and expression data to infer patient-specific pathway activities in TCGA GBM samples (6). In that report, the authors identified GBM subtypes using pathways inferred from the National Cancer Institute-Nature Pathway Interaction Database.

Here, we present a new approach to identify genes that tumors require for progression and survival, with patient-level specificity, by exhaustively and rapidly detecting correlations among gene expression and mutation. The method makes inferences directly from a collection of cancer genome samples and does not depend on pre-existing knowledge of gene function or interactions. We propose that this unbiased approach has utility to complement the findings of current gene set and pathway-based methods. We apply the method to examine the correlation between expression and mutation in TCGA GBM tumor samples. Our results suggest that this approach can be useful for identifying genes that participate in cancer progression, networks of genes that promote cancer via combined genetic and transcriptome alterations, druggable cancer-specific genes, and synthetic lethal interactions.

Materials and Methods

We developed a novel computational method to identify genes potentially important in tumorigenesis and cancer-specific survival genes from correlations among somatic mutation and expression in cancer genomics data (Fig. 1). The algorithm compares the sample (patient)-specific mutation status of each gene with the expression level of each gene, across all tumor samples. Genes with drastic mutation-correlated differential expression, and the corresponding mutated genes, are returned for analysis. The algorithm also identifies statistically significant mutation–mutation coincidence and mutual exclusivity. Gene networks are constructed containing all significant correlations and automated literature searches are used to illuminate clinically relevant findings. Findings presented here were identified using all TCGA GBM samples for which expression and mutation data were available and have a value of $P < 0.01$ and a false discovery rate (FDR) of less than 0.05.

Algorithm

We begin by building 2 matrices, 1 expression and 1 mutation, which are gene (row) by sample (column; Fig. 1A). At this stage, the expression matrix is populated by the factored, 3-platform data (see Data) and the mutation matrix is binary: 1 (true) if any mutation (see Data) occurs in a particular gene in a particular sample, otherwise the element is 0 (false).

Next, 2-class, unpaired Significance Analysis of Microarrays (SAM; ref. 17) is used to find genes that are differentially expressed with respect to the mutation status of a particular gene across all samples (i.e., the 2 classes are defined by the binary mutation vector for that particular gene from the mutation matrix; Fig. 1B). SAM was employed using a moderated *t*-statistic and the random seed was set to a constant (rand = 123) for reproducibility. To correct for multiple testing, 100 random permutations of the class labels were made and a cutoff FDR of 0.05 applied. Genes with an FDR of less than 0.05 are considered to have significant mutation-correlated differential expression and are passed to the next stage of the algorithm.

Next, an expression matrix is created, this time only containing genes deemed to have significant mutation-correlated differential expression in the previous step. Then, the matrix is converted to 2 binary matrices (1 for significant overexpression and 1 for significant underexpression) with the following calculation: (i) The *z*-score for each expression matrix element is calculated with respect to that element's row (i.e., gene specific); this is repeated for each row (gene). (ii) For the overexpressed binary matrix, any element with a *z*-score > 2.0 is 1 (true), otherwise the element is 0 (false); for the underexpressed binary matrix an element is 1 if the *z*-score < −2.0 and 0 otherwise (Fig. 1C). Then, Fisher's exact *P* value is calculated for each gene in the expression matrix by populating a 2 × 2 contingency table with a binary expression vector (category 1) and the mutation vector (category 2); this process is repeated for each binary expression vector from the binary expression matrix (Fig. 1D). This calculation allowed us to recover only genes that had drastic mutation-correlated over- and underexpression and to assign each correlation with an exact *P* value. Mutation-correlated over- and underexpressed genes with a value of $P < 0.01$ (Fisher's exact test) and an FDR of less than 0.05 (see Multiple Testing Correction) are hierarchically clustered using heatmap.2 from the R package, and mutations are plotted across the samples (Fig. 1E). The entire process is repeated once for each mutated gene.

Pairwise mutation–mutation correlation

Two-by-two contingency tables were constructed for every pairwise mutation vector to find significant ($P < 0.01$, Fisher's exact test) mutational co-occurrence and mutual exclusivity. Coincident pairwise mutation in at least 3 samples was additionally required to declare significant mutational co-occurrence.

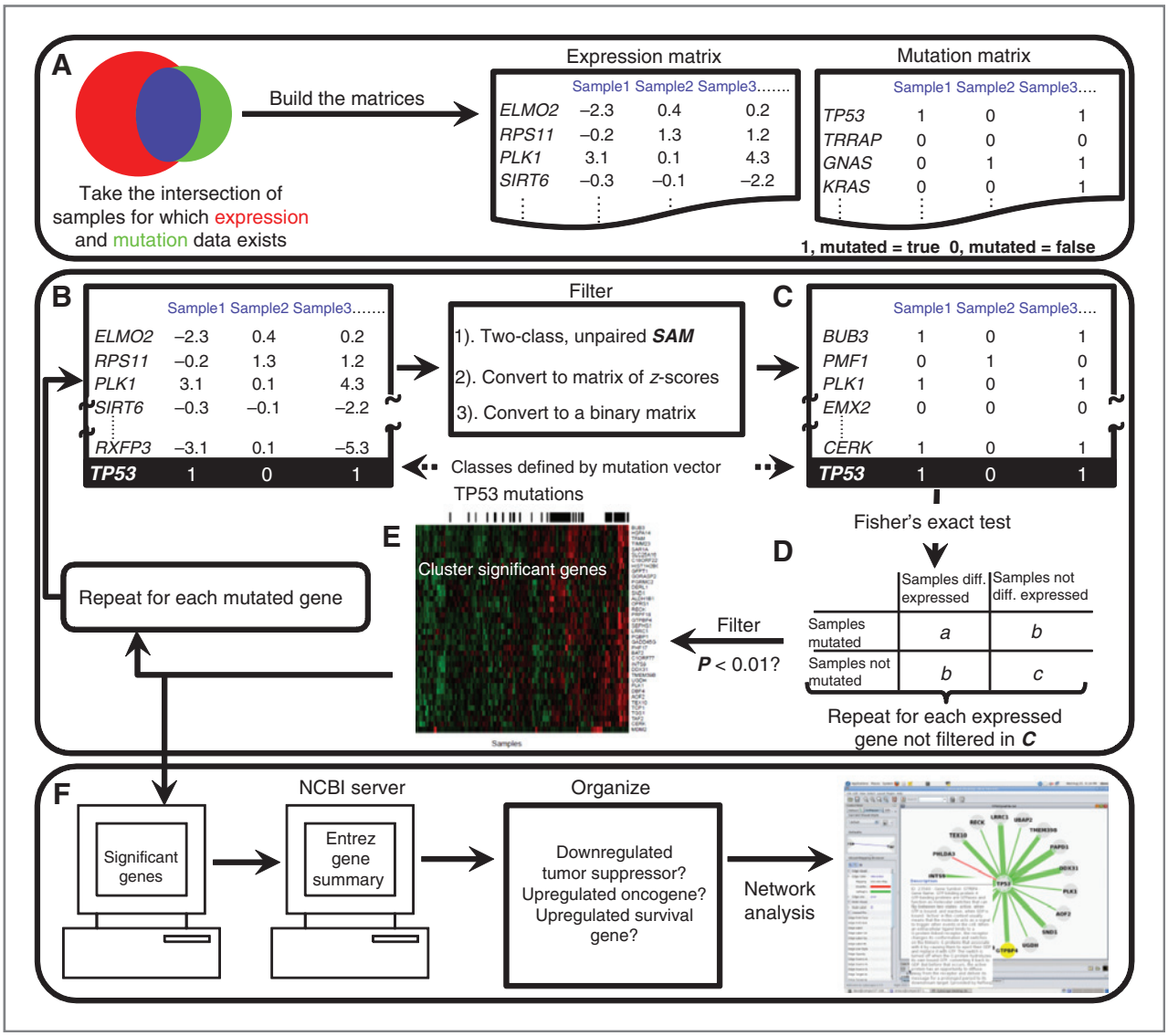

Figure 1 . Schematic overview of the algorithm developed for this study. A, expression and mutation matrices are built for samples containing both expression and sequence data. B–E, representative example of a single loop of the algorithm (toy example using TP53); there is 1 loop per mutated gene. B, a mutation vector (taken from the mutation matrix in A) is added to the expression matrix for each algorithmic loop (B–E). C, SAM is used to filter expressed genes (FDR < 0.05) using the mutation vector to define the classes (i.e., gene mutated in sample 1), and gene wild type in sample 0). Expressed genes passing the FDR cutoff point are then converted to a binary matrix by calculating the *z*-score at each element (with respect to that element's row) and applying a *z*-score cutoff point. D, genes in the binary expression matrix are each individually used to populate a 2×2 contingency table with the mutation vector defining the second category, and the exact *P* value is calculated and expressed genes are filtered again ($P < 0.01$, Fisher's exact test). E, genes passing (C) and (D) are hierarchically clustered and plotted with mutation status indicated for each sample. B–E are repeated for each mutated gene, each time starting in (B) with the complete expression matrix and a new mutation vector. F, Entrez Gene Summaries are obtained for all genes involved in significant mutation-correlated differential expression for subsequently identifying tumor suppressors and oncogenes in the data.

Multiple testing correction

For both mutation–mutation correlation and mutation-correlated over- and underexpression, a potential discovery is declared when Fisher's exact value of $P < 0.01$. For each potential discovery, the algorithm makes 1,000 random permutations of the columns (samples) and counts the correlations inferred from the permutated data (i.e., false discoveries). If the calculated FDR is greater than 0.05, the potential discovery is rejected as false. Every correlation presented in this article has a Fisher's exact value of $P < 0.01$ with an FDR less than 0.05.

Data

We obtained expression data for GBM samples at the TCGA web site (http://tcga-data.nci.nih.gov/docs/publications/gbm_exp/). These expression data were gathered on 3 individual microarray platforms, including Affymetrix Human Exon ST GeneChips, Affymetrix HT-HG-U133A GeneChips,

and custom-designed Agilent 244,000 feature gene expression microarrays (11). A single estimate of the relative expression for each gene in each sample was obtained using factor analysis (11). We removed any gene that had a missing value in any sample used for this study (reducing the total from 11,861 to 11,828).

We obtained phase I GBM sequence data from the TCGA web site (http://tcga-data.nci.nih.gov/tcga/). We obtained phase II GBM sequence data from Baylor College of Medicine (David A. Wheeler, personal communication). All mutations labeled *Validated* and *Nonsilent*, and *Somatic* or *LOH* were used. There were a total of 583 genes that met these criteria and 149 samples for which both expression and mutation data were available.

Literature mining

All literature mining used to highlight results (Table 1) was automated to increase efficiency and reduce user bias. Importantly, the literature mining was used only to interpret results, not as an input to the algorithm. To highlight potentially important genes that were identified, the algorithm searched 2,438,505 abstracts and titles for PubMed keyword "cancer" and the gene of interest; the same procedure was carried out for 16,237 GBM-specific articles. To determine whether genes had been described in previous studies using TCGA GBM data, we downloaded references 5 to 12 and converted them to text for automated searching. All mutation-correlated overexpressed genes were cross-referenced with a list of known GBM survival genes (Table 2; ref. 18). We retrieved summaries from the Entrez Gene database; these summaries were used to determine whether mutation-correlated overexpressed genes were oncogenes and whether mutation-correlated underexpressed genes were tumor suppressors (Fig. 1F and Table 3).

The algorithm developed for this study was written in Python (Fig. 1). Entrez PubMed and Gene Summary database queries were made using Biopython. All calls to R and Bioconductor were made via the R interface for Python, RPy2 (http://rpy.sourceforge.net/rpy2.html). For the TCGA GBM data set used here, total algorithm running time was less than 5 hours on a Linux workstation (2-core, 1.86-GHz processor, and 4 GB of RAM).

Supplementary data

Supplementary data for this article include 6 supplementary files: SYNE1 and IDH1 survival analysis and occurrence of mutations highlighted in this study in the COSMIC database (Supplementary Fig. S1 and Supplementary Table S1); individual heatmaps for each of the 41 genes whose mutation status is significantly correlated with the over- and underexpression of other genes (Supplementary Heatmaps); the raw data, including *P* values, for every correlation returned by our algorithm (Supplementary mutation–expression correlation and Supplementary mutation–mutation correlation); the sample-specific mutation type (e.g., nonsense, splice site, frame shift) for each mutated gene highlighted in the study, the zygosity, and the CHASM predictions for each missense mutation (Supplementary Mutation type, zygosity, and score); legends for all spreadsheets (Supplementary Spreadsheet legends).

Results and Discussion

Table 1 shows statistics for all genes where mutation status is significantly correlated with the drastic over- or underexpression, across tumor samples, of other genes. Our clustering scheme required genes be mutated in at least 2 samples, which reduced the total TCGA GBM set from 583 to 307. Forty-one of these mutated genes (~13%) were correlated with the drastic over- or underexpression of at least 2 of the 11,828 genes for which expression data were available. The low fraction of such correlations returned by our method partially reflects the stringency of the tests used to determine significance (see Materials and Methods). Comparing the numbers in columns 2 and 3 of Table 1 shows that there is no intrinsic bias of the algorithm to infer mutation-correlated over- or underexpression from frequency of mutation. For instance, HPN and IDH1 are each mutated in 11 samples, and IDH1 is correlated with the drastic over- or underexpression of 1,001 genes, whereas HPN is only correlated with the drastic over- or underexpression of 3 genes. Low-frequency mutations also show a distribution of correlated expression. In the case of mitogen-activated protein kinase (MAPK) 9, which was mutated in only 2 samples, there are 396 genes with correlated over- or underexpression. Conversely, CHL1 was mutated in 2 samples and only correlated with the differential expression of 2 genes.

If tumors select for genetic alterations that coordinate to promote cancer progression, then identifying coordinated genetic alterations could be useful to identify genes involved in tumorigenesis. Indeed, our approach identifies genes generally accepted to be involved in tumorigenesis (e.g., *ATM*, *FGFR1*, *IDH1*, *MET*, *MSH6*, *NF1*, *RB1*, and *TP53*). It is particularly difficult to assess the capacity of a genetic alteration to participate in cancer progression when that alteration is low frequency in the population; our approach identifies genes potentially involved in tumorigenesis that are mutated with low frequency in TCGA GBM tumor samples. For instance, ATM, KLF6, and LEMD3 are low-frequency mutations in TCGA GBM tumor samples and have completely overlapping comutation ($P = 9 \times 10^{-5}$ for each pairwise interaction, Fisher's exact test). And, these low-frequency mutations are each highly correlated with the drastic over- or underexpression of 165 other genes. These observations suggest that ATM, KLF6, and LEMD3 may cooperatively promote tumorigenesis in some TCGA GBM samples.

EP300 and FGFR4, FBXW7 and FURIN, and EP400 and FN1 are also each exclusively comutated in TCGA GBM samples (Table 1). These 4 exclusively comutated sets of genes comprise 9 of the 41 mutated genes identified in this study (~22%), which may be unexpected. One potential explanation for this finding is that the mutant pairs have a specific epistatic relationship that is distinct from any of the mutations in isolation. A factor complicating the interpretation of the exclusively comutated sets is the occurrence of the of the so-called *mutator phenotype*. Each gene in the exclusively comutated sets is mutated in samples that are of the mutator phenotype, marked by higher-than-average mutation rates owing to mutation in mismatch repair genes. With the

Table 1. The 41 genes whose mutation status is correlated with drastic over- and underexpression of other genes

Gene symbol	Mutations	Differentially expressed genes	PubMed hits for cancer	PubMed hits for glioblastoma	Previously highlighted in study using TCGA GBM data?
AKR1C3	16	5	82	4	Yes
ANK1	3	2	9	0	No
ATM[a]	2	165	2,175	20	Yes
KLF6[a]	2	165	103	6	No
LEMD3[a]	2	165	5	0	No
CHL1	2	2	15	0	No
ConsReg523	7	10	0	0	No
DST (dystonin)	12	14	14	0	No
EP300[b]	2	8	155	0	No
FGFR4[b]	2	8	139	4	No
EP400[c]	3	6	9	0	No
FN1[c]	3	6	58	2	No
EPHB4	3	149	104	0	No
FBXW7[d]	2	259	109	4	No
FURIN[d]	2	259	252	8	Yes
FGFR1	3	2	491	19	Yes
HPN (Hepsin)	11	3	82	1	No
IDH1	11	1,001	126	41	Yes
INHBC	3	2	5	0	No
KCNG1	3	8	0	0	No
LGALS3BP	3	4	54	1	No
LUM	3	8	6	0	No
MADD	10	15	18	0	No
MAPK9	2	396	2	0	No
MARK1	2	20	14	0	No
MET (c-met)	3	10	1,674	52	Yes
MKI67	27	2	7,866	218	No
MSH6	4	3	320	9	Yes
MYST4	3	2	11	0	No
NF1	19	10	1,916	41	Yes
NOS3	3	59	190	1	No
PI15	2	47	1	0	No
PTK2B	2	55	38	0	No
RB1	9	12	1,025	21	Yes
SRGAP1	2	5	5	0	No
STK36	3	2	4	0	No
SYNE1	10	543	5	0	No
TCF12	2	28	25	0	Yes
TP53	48	38	4,011	189	Yes
TRPM3	4	318	1	0	No
WISP1	2	2	39	0	No

NOTE: *Gene symbol* is the mutated gene, *mutations* indicates the number of samples harboring mutation in the mutated gene. *Differentially expressed genes* indicates the number of genes whose expression is correlated with the mutation status of the mutated gene. *PubMed hits for "cancer"* is the number of times the mutated gene appeared in the title or abstract for articles returned by search term "cancer" (2,438,505 total articles returned) and similarly for *PubMed hits for "glioblastoma"* (16,237 total articles returned). *Previously highlighted in a study using TCGA GBM data* indicates whether the mutated gene appeared anywhere in the text of references 5 to 12. [a,b,c,d]Groups of exclusively comutated. See Materials and Methods for complete details.

exception of LEMD3, EP400, and FN1, all genes in the comutated sets are well-studied cancer genes, and recurrence of mutations in these genes highlights them as potentially important in the progression of some gliomas. But, because these mutations were found in samples displaying the mutator phenotype, the possibility that some of them are passenger mutations has to be considered.

Columns 3 to 5 of Table 1 are derived from automated literature searches. Although automated literature mining can be prone to false positives, large disparities among and within rows of Table 1 can highlight potentially important genes that can be investigated manually. For instance, KFL6 was found in the title or abstract of 103 articles on cancer and 6 specifically on GBM; however, KFL6 is a low-frequency mutation in TCGA GBM tumor samples and has never been highlighted in a study of TCGA GBM data. Manual investigation of PubMed IDs returned by our method indicates that *KFL6* is a well-studied cancer gene (19, 20). KLF6 is a putative tumor suppressor that mediates growth inhibition by overexpression of the cell-cycle inhibitor CDKN1A (19). TCGA GBM samples contain at least 1 mutation in a previously reported KLF6 glioma mutation site (S77; ref. 20). Similarly, *EPHB4*, *FGFR4*, *FURIN*, and *NOS3* are all thought to be important in cancer progression; these genes are mutated with low frequency in TCGA GBM tumor samples and not highlighted in previous studies using TCGA GBM data.

TP53 network

Figure 2 is a heatmap of genes whose over- or underexpression is significantly correlated with TP53 mutation. TP53 mutations clusters in 2 main groups on the right half the heatmap, with a few smaller clusters and outliers located among the samples. Aside from MDM2 (bottom of Fig. 2), all genes in Figure 2 are overexpressed when TP53 is mutated. MDM2 is overexpressed when TP53 is wild type (i.e., MDM2 overexpression is mutually exclusive with TP53 mutation).

TP53 is a well-studied cancer gene, therefore method efficacy can be considered on the basis of its ability to capture known correlations. For instance, MDM2 is a negative reg-

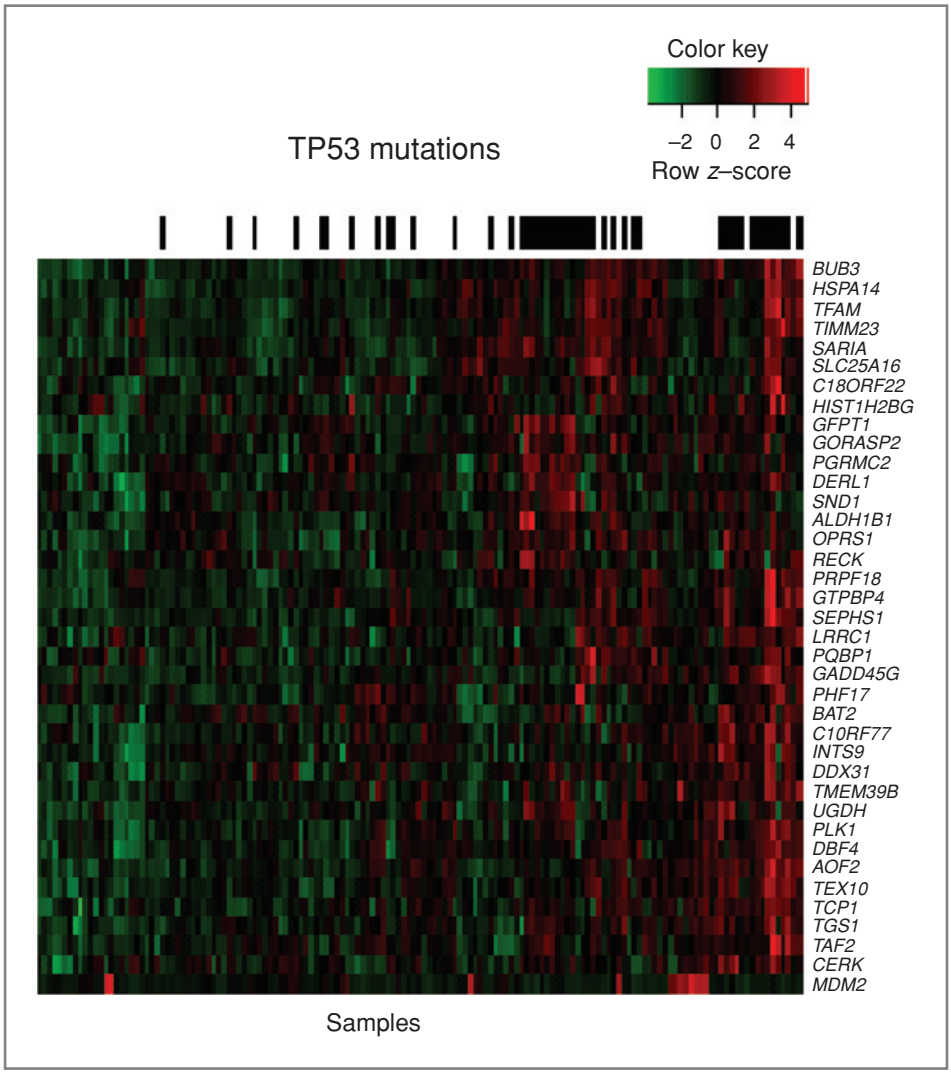

Figure 2. Hierarchical clustering of genes whose drastic over- and underexpression is correlated with TP53 mutation status. Black bars across the top of the heatmap indicate samples harboring TP53 mutation.

ulator of tumor suppressor TP53, therefore MDM2 overexpression and TP53 mutation can have a redundant phenotype and can be mutually exclusive (21); our method recovers this mutual exclusivity ($P = 0.0075$, Fisher's exact test).

The observation that PLK1 overexpression occurs in cancer cells harboring TP53 mutation led some groups to speculate that inhibition of PLK1 may specifically kill TP53 mutant cells (22, 23). Indeed, PLK1 inhibitors specifically kill cells harboring TP53 mutation (22, 23), suggesting TP53 and PLK1 may constitute a synthetic lethal interaction. We find a similar relation between TP53 and PLK1 in human TCGA GBM tumor samples ($P = 0.0099$, Fisher's exact test). Our method does not find significant correlation between PLK1 overexpression and the mutation status of any gene other than TP53.

DBF4 overexpression has been specifically linked to TP53 status (24). RNA-mediated interference of DBF4 was shown to specifically slow growth and reduce survival of melanoma cells (25). Our method found that DBF4 overexpression is correlated with TP53 mutation in TCGA GBM tumor samples ($P = 0.0099$, Fisher's exact test). TP53 and DBF4 may constitute a synthetic lethal pair, and DBF4 drugging might specifically kill cells harboring TP53 mutation. Our method does not find significant correlation between DBF4 overexpression and the mutation status of any gene other than TP53.

siRNA knockdown of the TP53-associated *TCP1* gene resulted in slowed growth in a ovarian carcinoma cell line (26). In a study using 186 breast cancer tumors, TCP1 subunit overexpression was shown to be correlated with *TP53* mutation (27). We find significant correlation ($P = 0.0099$, Fisher's exact test) between TCP1 overexpression and TP53 mutation in TCGA GBM tumor samples. *TP53* mutation may create a dependence on the overexpression of TCP1 and TCP1 may present a therapeutic vulnerability in some TP53-driven cancers.

BUB3 (28), *HSPA14* (*HSP60*; ref. 29), *TFAM* (30), *GFTP1* (*GFAT*; ref. 31), *DERL1* (32), *SND1* (33), *ALDH1B1* (34), *RECK* (35), *UGHD* (36), *AOF2* (*LSD1*; ref. 37), *GADD45G* (38), and *CERK* (ceramide kinase; ref. 39) have also been central genes in at least one cancer study where each was found to be overexpressed in certain cancers. DERL1 overexpression inspired Ran and colleagues (32) to target DERL1 with anti-DERL1 antibodies, which resulted in tumor growth suppression in mice. The UGDH inhibitors gallic acid and quercetin have strong antiproliferative effects in breast cancers overexpressing UGDH (36). siRNA-mediated knockdown of AOF2 (LSD1) slows neuroblastoma cell growth in cells overexpressing AOF2 (37). Repression of CERK in a human adenocarcinoma cell line overexpressing CERK induced apoptosis (40). HSPA14 (HSP60) inhibition can selectively induce apoptosis in tumor cells overexpressing HSPA14 (29). To the best of our knowledge, this is the first time, the overexpression of these genes has been linked to TP53 mutation. Because inhibition of these genes induces effects specific to cancer cells, they may be druggable targets in cancers mutated in TP53.

Subclusters in Figure 2 arise from tumor samples sharing genes with similar TP53 mutation–correlated expression. For example, PLK1, DBF4, AOF2, TCP1, and CERK (defined here as cluster *A*) cluster together because they have similar expression across all samples. Overexpression of each cluster A gene is associated with a druggable dependence in cancer cells (22, 23, 25, 26, 37, 40), and PLK1 (22), DBF4 (24), and TCP1 (27) are known to be overexpressed specifically in the context of TP53 mutation. Our method identifies groups of tumors that may have a dependence on multiple druggable targets. Tumor dependence on multiple overexpressed druggable genes may be of therapeutic relevance because low-concentration inhibitor cocktails could replace single-agent targeted therapies, resulting in increased therapeutic index (41).

Mutation-correlated differential expression among subclusters may also inform therapeutic regimens. For instance, GFPT1, GORASP2, PGRMC2, DERL1, SND1, ALDH1B1, OPRS1, and RECK (defined here as cluster *B*) are overexpressed in tumors distinct from those with cluster A gene overexpression. Because cluster A gene overexpression is a signature of cluster A gene dependence, cluster A gene inhibitors might inhibit tumors overexpressing cluster A genes more than tumors lacking cluster A gene overexpression. In that case, patients with cluster A signatures and patients with cluster B signatures may benefit from different drugging protocols, which our method highlights.

IDH1/SYNE1 networks

In a landmark study, Parsons and colleagues discovered a novel, high-frequency driver mutation in IDH1, highlighting the utility of unbiased genomics experiments (4). Focused studies by many groups confirmed the importance of IDH1 mutation in GBM. Of the 41 mutated genes returned by our method, IDH1 is one of the most studied genes in GBM (Table 1, column 5), which is striking considering its importance in GBM is recently discovered. Our method finds 1,001 genes have drastic over- or underexpression associated with IDH1 mutation status; this IDH1 network is by far the largest network returned by our method (Table 1, column 3). This suggests that IDH1 mutation is associated with an unique GBM (epi)genotype. Indeed, IDH1 mutation is a defining characteristic of the proneural GBM subtype (11) and the glioma CpG island methylator phenotype (12).

Figure 3 is a graph representation of all IDH1 nearest and second-nearest neighbors. In this graph, nodes represent mutated genes, overexpressed oncogenes or GBM survival genes, or underexpressed tumor suppressors returned by our method. These types of coordinated (de)activation can drive cancer, and second-nearest neighbors highlight networks connected by common genes.

We find that TCGA GBM tumors with IDH1 mutation are significantly correlated with the drastic overexpression of several known GBM survival genes (Fig. 3 and Table 2; ref. 18): *MPHOSPH1*, *POLR2F*, *ARHGEF11*, and *AKT3* ($P = 4.1 \times 10^{-5}$, 2.8×10^{-3}, 2.8×10^{-3}, and 2.8×10^{-3}, respectively). Because IDH1 mutation is a defining characteristic of specific GBM (epi)genotypes, druggable dependencies associated with IDH1 mutation status could be clinically relevant. M phase phosphoprotein 1 (MPHOSPH1) is known to be overexpressed in some bladder cancers (42). Recently, phase I/II trials using MPHOSPH1 peptide epitopes were shown to induce specific cytotoxic T lymphocytes against

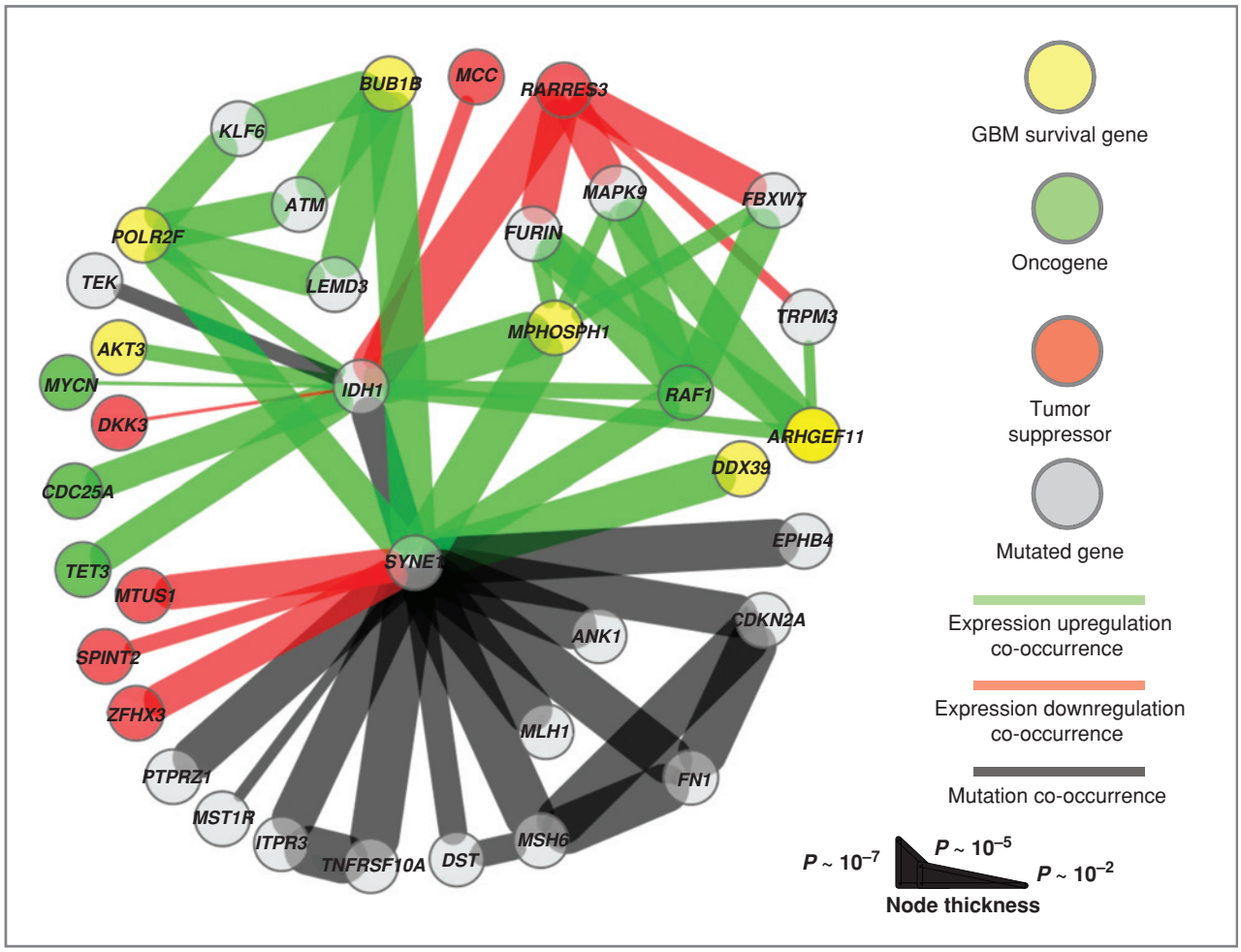

Figure 3. Nearest neighbors and second-nearest neighbors of IDH1. A graph representation of all nearest neighbor and second-nearest neighbor nodes comprising known GBM survival genes and oncogenes that were overexpressed, tumor suppressors that were underexpressed, and mutated genes.

bladder cancers overexpressing MPHOSPH1 (42). Our results suggest that similar approaches may benefit some gliomas mutated in IDH1. AKT3 overexpression was found in a significant fraction of breast and prostate cancers (43) and has been reported as a possible oncogene and a potential glioma survival gene (18). Indeed, oncogene addiction could be considered a type of survival gene dependence. Importantly, *AKT3* is a well-studied cancer gene, and inhibitors of AKT3 and other genes in the AKT3 pathway exist (43). Here, we find that drastic overexpression of AKT3 is exclusively and significantly associated with IDH1 mutation. Overexpression of the GBM survival genes *POLR2F* and *ARHGEF11* are known markers in colon (44) and gall bladder (45) cancer, respectively; unfortunately, we know of no drugs that target these genes. All known GBM survival genes whose drastic overexpression is correlated with mutation status of another gene, and the corresponding mutated genes, are shown in Table 2.

Our method identified several mutation-correlated overexpressed oncogenes and underexpressed tumor suppressors, the majority being significantly associated with IDH1 mutation status (Fig. 3 and Table 3). The underexpression of tumor suppressors RARRES3, DKK3, and MCC was significantly correlated with IDH1 mutation ($P = 8.0 \times 10^{-5}$, 9.0×10^{-3}, and 2.8×10^{-3}, respectively); DKK3 and MCC underexpression was exclusively associated with IDH1 mutation. Overexpression of the oncogenes *RAF1*, *MYCN*, *TET3*, and *CDC25A* was significantly correlated with IDH1 mutation ($P = 2.8 \times 10^{-3}$, 9.0×10^{-3}, 1.2×10^{-3}, and 1.2×10^{-3}, respectively); *MYCN*, *TET3*, and *CDC25A* overexpression was exclusively associated with IDH1 mutation. All known oncogenes whose drastic overexpression, and tumor suppressors whose drastic underexpression is correlated with mutation status of another gene, and the corresponding mutated genes, are shown in Table 3.

Our method finds that 543 genes have drastic over- or underexpression associated with SYNE1 mutation status; this SYNE1 network is the second largest network returned by our method (Table 1, column 3). Similarly, SYNE1 participates in significant mutational co-occurrence more than any other gene; there are 12 mutation–mutation interactions involving SYNE1 (Fig. 3). Also, *SYNE1* is the only gene with which MSH6

Table 2. Known GBM survival genes that are overexpressed in correlation with the mutation status of specific TCGA GBM genes

Mutated gene(s)	GBM survival gene	Known medical relevance
ATM, *LEMD3*, *KLF6*, *IDH1*, *SYNE1*	*MPHOSPH1*	Phase II epitope peptide vaccine
FBXW7, *FURIN*, *IDH1*, *MAPK9*, *SYNE1*	*POLR2F*	Prognostic marker in colon cancer
FBXW7, *FURIN*, *IDH1*, *MAPK9*, *SYNE1*, *TRPM3*	*ARHGEF11*	Marker in gall bladder cancer
ATM, *KLF6*, *LEMD3*, *SYNE1*	*BUB1B*	Aurora B inhibition by Hesperadin can prevent BUB1B kinetochore localization
IDH1	*AKT3*	Many inhibitors and upstream inhibitors
SYNE1	*DDX39*	Marker in several cancers

NOTE: *Mutated gene(s)* whose mutation status was correlated with the overexpression of a known *GBM survival gene* and *Known medical relevance* of the survival gene. Genes that are grouped by a rectangle were exclusively comutated.

and MLH1 have complete mutational overlap (1.1×10^{-5} and 2.2×10^{-4}, respectively); *MSH6* and *MLH1* are mismatch repair genes whose mutation is known to cause the so-called *mutator phenotype* in GBM (46).

SYNE1 mutation is high frequency in TCGA GBM tumor samples but has not been highlighted in previous studies using TCGA GBM data (Table 1). Similarly, our method does not find any previous correlation between GBM and SYNE1 mutation in the literature (Table 1). SYNE1 mutation is known to influence cerebellar ataxia and has recently been associated with lung, ovarian, and colorectal cancers (47). Our results suggest that SYNE1 mutation is important in TCGA GBM tumor samples and may be important in some glioblastomas in general.

We find that SYNE1 mutation is significantly correlated with the overexpression of several known GBM survival genes (Table 2). *BUB1B* is a chromosome instability gene known to be involved in cancer (48). The aurora B inhibitor hesperadin can prevent kinetochore localization of BUB1B and arrest cell-cycle progression (49); hesperadin has not yet been proven effective in cancer clinical trials. We find this known GBM survival and chromosome instability gene to be overexpressed in the presence of SYNE1 mutation ($P = 8.6 \times 10^{-4}$), suggesting BUB1B as a potential therapeutic target in some SYNE1-mutated gliomas. *DDX39* is known to be overexpressed in several cancer types (50) and is a known GBM survival gene (18). We suggest that there is possible a connection between these results, in that DDX39 dependency may present as DDX39 overexpression. And, this dependency is significantly correlated with the mutation status of SYNE1 in TCGA GBM samples ($P = 8.6 \times 10^{-4}$). Other survival genes having overexpression significantly correlated with SYNE1 mutation status include MPHOSPH1 and POLR2F ($P = 6.5 \times 10^{-4}$ and 2.1×10^{-3}, respectively) and were described above. Our method also finds that the underexpression of the MTUS1, ZFHX3, and SPINT2 tumor suppressors is significantly and exclusively correlated with the mutation status of SYNE1 ($P = 2.1 \times 10^{-3}$, 2.1×10^{-3}, and 4.1×10^{-3}, respectively). *RAF1* oncogene overexpression is significantly correlated with SYNE1 mutation status ($P = 2.1 \times 10^{-3}$).

Other considerations

One important distinction to make, when considering alteration co-occurrence in cancer, is whether identified interactions have true cellular dependence or whether they are correlated for an unidentified reason. For instance, our

Table 3. Mutation-correlated inactivation of tumor suppressors and activation of oncogenes (as inferred from expression)

Mutated gene(s)	Underexpressed tumor suppressor	Overexpressed oncogene
FBXW7	RARRES3	*RAF1*
FURRIN		
IDH1	RARRES3, DKK3, MCC	*RAF1*, *MYCN*, *TET3*, *CDC25A*
MAPK9	RARRES3	*RAF1*
TRPM3	RARRES3, DRAM	*KRAS*, *CRKL*
SYNE1	MTUS1, ZFHX3, SPINT2	*RAF1*

NOTE: *Mutated gene(s)* whose mutation status is correlated with *underexpressed tumor suppressors* and *overexpressed oncogenes*. Genes that were exclusively comutated are grouped into single rows [*mutated gene(s)*].

method recovered several GBM survival genes whose overexpression was correlated with IDH1 mutation status. But, IDH1 mutation was found to be associated with a broadly altered (epi)genotype in this and other GBM studies. Therefore, IDH1 mutation and survival gene overexpression could be selected for by similar or overlapping hubs from the *IDH1 network*, but not by each other. In that scenario, the complex networks could vary among patients and cancer types reducing the generalizability of drugging protocols. Furthermore, any inhibitor targeting an overexpressed gene will be limited in efficacy to scenarios where that gene is significantly overexpressed in the patients tumor relative to their healthy tissue.

Elucidating true interaction dependence could also be informative. For instance, if mutation in hypothetical gene *A* created a strict cellular dependence on the overexpression of hypothetical gene *B*, then by definition, there would be a requirement for gene *B* overexpression to precede gene *A* mutation during cancer progression. This temporal ordering would be required because cells harboring mutation in gene *A*, but not overexpressing gene *B*, would be eliminated from the population. In cancer genomics data, this would manifest as significant correlation between gene *A* mutation and gene *B* overexpression, and on an average, a greater number of samples overexpressing gene *B*, compared with those harboring gene *A* mutation. The requirement for such temporal ordering could be exploited for prognosis as well as provide an obvious therapeutic target.

Conclusions

In this report, we developed an intuitive and unbiased method to exhaustively interrogate cancer genomics data to identify genes that tumors require for progression and survival. The method identified many genes known to promote GBM pathogenesis and highlighted several genes not previously associated with GBM as potentially important in GBM pathogenesis. In addition, the algorithm identified known druggable cancer-specific dependencies, survival genes, and potential synthetic lethal interactions. And, all observations were identified with patient specificity, which could increase clinical utility.

This algorithm should be a useful complement to existing methods. Because it is exhaustive, and unbiased in that all genes are tested regardless of prior association to disease, our new algorithm may identify novel correlations that add to the existing/emerging picture of gliomas and caner in general. Furthermore, development of model-free approaches, such as those developed in this study, may be applicable to a wide range of genes and pathways as they do not rely on previously curated pathway or interaction databases.

A useful addition to our algorithm might be to consider site- or domain-specific mutation. Although this is expected to be noisy for most genes, genes with multiple domain-specific functions may influence distinct, mutation-specific regulatory changes. One difficulty in implementing such a strategy would be distinguishing protein functional regions in an automated fashion.

One improvement to our method would be the ability to automatically return known inhibitors for inferred therapeutic vulnerabilities. It is not immediately obvious how this improvement could be implemented, owing to a lack of systematic annotation in the literature; however, assembling the correct drug databases might be one approach. If successful, clinical cancer genomics data would be algorithmic input, and the output could consist of therapeutic vulnerabilities ranked by known druggability.

Important open questions include the origin of drug resistance and the generalizability of synthetic lethal interactions. Most inhibitors targeting a specific driver gene have only modest success, often owing to off-target alterations. Similarly, synthetic lethal killing of tumor cells with generalizability has yet to be shown, suggesting the potential existence of a *synthetic lethal network*. Therefore, a comprehensive list of compensatory alterations that cause drug resistance or facilitate viability in the presence of targeted synthetic lethality may be useful. The information imparted from such a compendium could allow clinicians to *cut cancer off at the pass*. To that end, the combined effort of high-throughput cancer (epi) genomics experiments and complementary bioinformatics approaches is indispensable.

Disclosure of Potential Conflicts of Interest

Under agreements between the Johns Hopkins University and Agios Pharmaceuticals, R. Karchin is entitled to a share of the royalties received by the University on sales of products related to IDH genes. D.L. Masica declared no potential conflicts of interest.

Acknowledgment

The authors thank Dr. Bert Vogelstein for his critical reading of the manuscript.

Grant Support

This work was funded by NIH National Cancer Institute grant CA135877 and NSF DBI CAREER award 0845275 to R. Karchin.

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

Received January 20, 2011; revised April 13, 2011; accepted May 1, 2011; published OnlineFirst May 9, 2011.

References

1. Jones S, Zhang X, Parsons DW, Lin JC-H, Leary RJ, Angenendt P, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science 2008;321: 1801–6.

2. McLendon R, Friedman A, Bigner D, Van Meir E, Brat D, Mastrogianakis G, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 2008; 455: 1061–8.

3. Yeang C-H, McCormick F, Levine A. Combinatorial patterns of somatic gene mutations in cancer. FASEB J 2008;22:2605–22.
4. Parsons DW, Jones S, Zhang X, Lin JC-H, Leary RJ, Angenendt P, et al. An integrated genomic analysis of human glioblastoma multiforme. Science 2008;321:1807–12.
5. Gaire RK, Bailey J, Bearfoot J, Campbell IG, Stuckey PJ, Haviv I. MIRAGAA—a methodology for finding coordinated effects of microRNA expression changes and genome aberrations in cancer. Bioinformatics 2010;26:161–7.
6. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics 2010;26:237–45.
7. Brennan C, Momota H, Hambardzumyan D, Ozawa T, Tandon A, Pedraza A, et al. Glioblastoma subclasses can be defined by activity among signal transduction pathways and associated genomic alterations. PLoS One 2009;4:e7752.
8. Freire P, Vilela M, Deus H, Kim Y-W, Koul D, Colman H, et al. Exploratory analysis of the copy number alterations in glioblastoma multiforme. PLoS One 2008;3:e4076.
9. Cerami E, Demir E, Schultz N, Taylor BS, Sander C. Automated network analysis identifies core pathways in glioblastoma. PLoS One 2010;5:e8918.
10. Bredel M, Scholtens DM, Harsh GR, Bredel C, Chandler JP, Renfrow JJ, et al. A network model of a cooperative genetic landscape in brain tumors. JAMA 2009;302:261–75.
11. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 2010;17:98–110.
12. Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, et al. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell 2010;17:510–22.
13. Luo J, Solimini NL, Elledge SJ. Principles of cancer therapy: oncogene and non-oncogene Addiction. Cell 2009;136:823–37.
14. McManus KJ, Barrett IJ, Nouhi Y, Hieter P. Specific synthetic lethal killing of RAD54B-deficient human colorectal cancer cells by FEN1 silencing. Proc Natl Acad Sci U S A 2009;106:3276–81.
15. Ikediobi ON. Somatic pharmacogenomics in cancer. Pharmacogenomics J 2008;8:305–14.
16. Singh A, Settleman J. Oncogenic K-ras "addiction" and synthetic lethality. Cell Cycle 2009;8:2676–8.
17. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001;98:5116–21.
18. Thaker NG, Zhang F, McDonald PR, Shun TY, Lewen MD, Pollack IF, et al. Identification of survival genes in human glioblastoma cells by small interfering RNA screening. Mol Pharmacol 2009;76:1246–55.
19. Narla G, Heath KE, Reeves HL, Li D, Giono LE, Kimmelman AC, et al. KLF6, a candidate tumor suppressor gene mutated in prostate cancer. Science 2001;294:2563–6.
20. Jeng Y-M, Hsu H-C. KLF6, a putative tumor suppressor gene, is mutated in astrocytic gliomas. Int J Cancer 2003;105:625–9.
21. Ichimura K, Bolin MB, Goike HM, Schmidt EE, Moshref A, Collins VP. Deregulation of the p14ARF/MDM2/p53 pathway is a prerequisite for human astrocytic gliomas with G1-S transition control gene abnormalities. Cancer Res 2000;60:417–24.
22. Degenhardt Y, Greshock J, Laquerre S, Gilmartin AG, Jing J, Richter M, et al. Sensitivity of cancer cells to Plk1 inhibitor GSK461364A is associated with loss of p53 function and chromosome instability. Mol Cancer Ther 2010;9:2079–89.
23. Sur S, Pagliarini R, Bunz F, Rago C, Diaz LA, Kinzler KW, et al. A panel of isogenic human cancer cells suggests a therapeutic approach for cancers with inactivated p53. Proc Natl Acad Sci U S A 2009;106:3964–9.
24. Bonte D, Lindvall C, Liu H, Dykema K, Furge K, Weinreich MCdc7-Dbf4 kinase overexpression in multiple cancers and tumor cell lines is correlated with p53 inactivationNeoplasia 2008;10:920–31.
25. Nambiar S, Mirmohammadsadegh A, Hassan M, Mota R, Marini A, Alaoui A, et al. Identification and functional characterization of ASK/Dbf4, a novel cell survival gene in cutaneous melanoma with prognostic relevance. Carcinogenesis 2007;28:2501–10.
26. Macleod K, Mullen P, Sewell J, Rabiasz G, Lawrie S, Miller E, et al. Altered ErbB receptor signaling and gene expression in cisplatin-resistant ovarian cancer. Cancer Res 2005;65:6789–800.
27. Ooe A, Kato K, Noguchi S. Possible involvement of CCT5, RGS3, and YKT6 genes up-regulated in p53-mutated tumors in resistance to docetaxel in human breast cancers. Breast Cancer Res Treat 2007;101:305–15.
28. Yuan B, Xu Y, Woo J-H, Wang Y, Bae YK, Yoon D-S, et al. Increased expression of mitotic checkpoint genes in breast cancer cells with chromosomal instability. Clin Cancer Res 2006;12:405–10.
29. Ghosh JC, Dohi T, Kang BH, Altieri DC. Hsp60 regulation of tumor cell apoptosis. J Biol Chem 2008;283:5188–94.
30. Cormio A, Guerra F, Cormio G, Pesce V, Fracasso F, Loizzi V, et al. The PGC-1[alpha]-dependent pathway of mitochondrial biogenesis is upregulated in type I endometrial cancer. Biochem Biophys Res Commun 2009;390:1182–5.
31. Paterson A, Kudlow J. Regulation of glutamine:fructose-6-phosphate amidotransferase gene transcription by epidermal growth factor and glucose. Endocrinology 1995;136:2809–16.
32. Ran Y, Hu H, Hu D, Zhou Z, Sun Y, Yu L, et al. Derlin-1 is overexpressed on the tumor cell surface and enables antibody-mediated tumor targeting therapy. Clin Cancer Res 2008;14:6538–45.
33. Ho J, Kong J-W-F, Choong L-Y, Loh M-C-S, Toy W, Chong P-K, et al. Novel breast cancer metastasis-associated proteins. J Proteome Res 2008;8:583–94.
34. The Gene expression profiles of medulloblastoma cell lines resistant to preactivated cyclophosphamide. Curr Cancer Drug Targets 2008;8:172–9.
35. Kitajima S, Miki T, Takegami Y, Kido Y, Noda M, Hara E, et al. Reversion-inducing cysteine-rich protein with Kazal motifs interferes with epidermal growth factor receptor signaling. Oncogene 2011;30:737–50.
36. Hwang EY, Huh J-W, Choi M-M, Choi SY, Hong H-N, Cho S-W. Inhibitory effects of gallic acid and quercetin on UDP-glucose dehydrogenase activity. FEBS Lett 2008;582:3793–7.
37. Schulte JH, Lim S, Schramm A, Friedrichs N, Koster J, Versteeg R, et al. Lysine-specific demethylase 1 is strongly expressed in poorly differentiated neuroblastoma: implications for therapy. Cancer Res 2009;69:2065–71.
38. Flores O, Burnstein KL. GADD45{gamma}: a new vitamin D-regulated gene that is antiproliferative in prostate cancer cells. Endocrinology 2010;151:4654–64.
39. Ruckhäberle E, Karn T, Rody A, Hanker L, Gätje R, Metzler D, et al. Gene expression of ceramide kinase, galactosyl ceramide synthase and ganglioside GD3 synthase is associated with prognosis in breast cancer. J Cancer Res Clin Oncol 2009;135:1005–13.
40. Mitra P, Maceyka M, Payne SG, Lamour N, Milstien S, Chalfant CE, et al. Ceramide kinase regulates growth and survival of A549 human lung adenocarcinoma cells. FEBS Lett 2007;581:735–40.
41. Teicher BA. Combinations of PARP, hedgehog and HDAC inhibitors with standard drugs. Current Opin Pharmac 2010;10:397–404.
42. Obara W, Tsunoda T, Yoshida K, Kanehira M, Takata R, Katagiri T, et al. Phase I/II study of novel HLA-A24 restricted DEPDC1 and MPHOSPH1 peptide vaccine for bladder cancer. In: J Clin Oncol (Meeting Abstr) 2010;28:e13122.
43. Lindsley C, Barnett S, Layton M, Bilodeau M. The PI3K/Akt pathway: recent progress in the development of ATP-competitive and allosteric Akt kinase inhibitors. Curr Cancer Drug Targets 2008;8:7–18.
44. Antonacopoulou AG, Grivas PD, Skarlas L, Kalofonos M, Scopa CD, Kalofonos HP. POLR2F, ATP6V0A1 and PRNP expression in colorectal cancer: new molecules with prognostic significance?Anticancer Res 2008;28:1221–7.
45. Kim J, Kim H, Lee K, Lee J, Choi S, Paik S, et al. Gene expression profiles in gallbladder cancer: the close genetic similarity seen for early and advanced gallbladder cancers may explain the poor prognosis. Tumor Biol 2008;29:41–9.

46. Purow B, Schiff D. Advances in the genetics of glioblastoma: are we reaching critical mass?Nat Rev Neurol 2009;5:419–26.
47. Doherty JA, Rossing MA, Cushing-Haugen KL, Chen C, Van Den Berg DJ, Wu AH, et al. ESR1/SYNE1 polymorphism and invasive epithelial ovarian cancer risk: an Ovarian Cancer Association Consortium Study. Cancer Epidemiol Biomarkers Prev 2010;19:245–50.
48. Ricke RM, van Ree JH, van Deursen JM. Whole chromosome instability and cancer: a complex relationship. Trends Genet 2008;24:457–66.
49. Hauf S, Cole RW, LaTerra S, Zimmer C, Schnapp G, Walter R, et al. The small molecule Hesperadin reveals a role for Aurora B in correcting kinetochore–microtubule attachment and in maintaining the spindle assembly checkpoint. J Cell Biol 2003;161:281–94.
50. Sugiura T, Nagano Y, Noguchi Y. DDX39, upregulated in lung squamous cell cancer, displays RNA helicase activities and promotes cancer cell growth. Cancer Biol Ther 2007;6:957.