Identification of rare cancer driver mutations by network reconstruction

Ali Torkamani and Nicholas J. Schork

Genome Res. 2009 19: 1570-1578 originally published online July 2, 2009
Access the most recent version at doi:10.1101/gr.092833.109

Identification of rare cancer driver mutations by network reconstruction

Ali Torkamani and Nicholas J. Schork[1]

The Scripps Translational Science Institute and Scripps Genomic Medicine, Scripps Health and The Scripps Research Institute, La Jolla, California 92037, USA

Recent large-scale tumor resequencing studies have identified a number of mutations that might be involved in tumorigenesis. Analysis of the frequency of specific mutations across different tumors has been able to identify some, but not all of the mutated genes that contribute to tumor initiation and progression. One reason for this is that other functionally important genes are likely to be mutated more rarely and only in specific contexts. Thus, for example, mutation in one member of a collection of functionally related genes may result in the same net effect, and/or mutations in certain genes may be observed less frequently if they play functional roles in later stages of tumor development, such as metastasis. We modified and applied a network reconstruction and coexpression module identification-based approach to identify functionally related gene modules targeted by somatic mutations in cancer. This method was applied to available breast cancer, colorectal cancer, and glioblastoma sequence data, and identified Wnt / TGF-beta cross-talk, Wnt / VEGF signaling, and MAPK / focal adhesion kinase pathways as targets of rare driver mutations in breast, colorectal cancer, and glioblastoma, respectively. These mutations do not appear to alter genes that play a central role in these pathways, but rather contribute to a more refined shaping or "tuning' of the functioning of these pathways in such a way as to result in the inhibition of their tumor-suppressive signaling arms, and thereby conserve or enhance tumor-promoting processes.

[Supplemental material is available online at http://www.genome.org.]

A number of recent tumor resequencing studies have been pursued to identify mutations that are likely to cause tumor formation (Wood et al. 2007; Cancer Genome Atlas Research Network 2008; Ding et al. 2008; Jones et al. 2008b; Parsons et al. 2008). Mutations that are likely to contribute to or cause tumorigenesis are termed "driver" mutations, and those likely nonfunctional or neutral mutations that simply build up during the unchecked cell turnover and proliferation that is the hallmark of tumor formation are termed "passenger" mutations (Torkamani et al. 2008). Identifying driver mutations and distinguishing them from mere passenger mutations is not trivial. Many, if not most studies seeking to identify driver mutations exploit strategies that consider the frequency of mutations as the sole criterion for differentiating driver mutations from passenger mutations—the intuition being that mutations in genes observed more often across a set of tumors are more likely to have resulted from a cancer cell selection process and, hence, represent mutations essential for tumorigenesis (Torkamani et al. 2008). However, resequencing studies suggest that rare mutations are likely to make up the vast majority of mutations contributing to tumorigenesis. Identifying such mutations poses a number of challenges and is the focus of this work.

As an example of an attempt to identify driver mutations using a frequency-based strategy, a recent resequencing study of 11 breast and 11 colorectal tumor samples identified mutations in 1138 genes among the breast cancer samples and 849 genes among the colorectal cancer samples (Wood et al. 2007). The genes that exhibited frequent mutations were then resequenced in an additional set of tumors to identify candidate cancer genes, termed "CAN-genes" by the investigators. However, given estimates of the number of genes involved in tumorigenesis based upon positive selection for nonsynonymous mutations, a number of mutated genes may have carried functional mutations and contributed to tumorigenesis, but were indistinguishable from background mutations based on their mutation frequency. Given the wide variety of mutations likely to be involved in tumorigenesis, frequency-based approaches for identification of rare cancer driver mutations are not likely to be successful without enormous investments in sample collection and characterization in order to achieve adequate power. Thus, methods for identification of rare driver mutations are in need.

Previous efforts to detect rare driver mutations have focused on known pathways or known direct interactions between mutated genes, resulting in descriptions of tumorigenic processes in very general terms, and hence, lack specificity with respect to the role of specific mutations in the tumorigenic process (Hernández et al. 2007; Lin et al. 2007). In this study, we applied a network reconstruction and gene coexpression module-based approach to identify distinct coexpression modules containing a larger number of mutated genes than expected by chance. This approach is a modification and application of the general framework for weighted gene coexpression network analysis described by Zhang and Horvath (2005), Horvath et al. (2006), and Oldham et al. (2006). This unbiased approach does not rely on prior knowledge of the biological relationships between genes, but rather attempts to reconstruct sets of coordinately acting genes in order to define, de novo, biological processes affected by cancer mutations. We have developed and applied this approach to genes known to be mutated in breast cancer, colorectal cancer, and glioblastoma in order to identify groups of genes likely to bear functionally important driver mutations.

We find that the resultant coexpression modules bearing an excess of somatic mutations are likely to alter signaling pathways known to be important for late tumorigenesis, such as cross-talk of the Wnt and TGF-beta signaling pathways in breast cancer, the Wnt and VEGF paracrine and autocrine signaling pathways in colorectal cancer, and MAPK and focal adhesion kinase signaling pathways in glioblastoma. These mutations generally do not affect

[1] **Corresponding author.**
E-mail nschork@scripps.edu; fax (858) 546-9284.

Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.092833.109.

19:1570–1578 © 2009 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/09; www.genome.org

the primary signaling members of the pathways (i.e., receptors and their immediate signaling partners) and would not likely be characterized as important mutations within these pathways using classical pathway analysis approaches. We speculate that these mutations "fine tune" distinct signaling arms of each pathway, such that the tumor-suppressive activities of these pathways are diminished, while tumor promoting activities enabled by these mutations are favored (or simply left intact). Ultimately, our "systems biology" approach to identifying rare mutations contributing to tumorigenesis appears to have promise.

Results

Cancer network reconstruction

To identify gene coexpression modules targeted by somatic mutations in the various cancer types, we first reconstructed breast, colorectal, and glial normal and cancerous tissue gene coexpression networks. We chose to use the ARACNE algorithm for this purpose because of its proven superior performance over other algorithms and its computational feasibility on a whole-genome scale (Margolin et al. 2006). Although there are many algorithms and strategies for reconstructing gene coexpression networks, including those that exploit a simple pairwise correlation matrix of gene-expression levels and clustering algorithms, the mutual information approach in the ARACNE algorithms has been shown to provide superior results in related contexts (Priness et al. 2007). Other investigators have applied the mutual information approach to related gene-clustering approaches previously (Butte and Kohane 2000; Daub and Sonnhammer 2008). Importantly, mutual information more accurately captures nonlinear expression relationships between genes. To demonstrate this, we compared the mutual information scores vs. the absolute value of the Pearson's correlation coefficients derived from the breast cancer expression datasets. Figure 1 presents the Pearson's correlation versus mutual information for genes in breast module 26 with all other genes (Fig. 1A) or only with other genes from breast module 26 (Fig. 1B). The importance of breast module 26 is described later on. In general, while Pearson's correlation captures the relationship between module genes well, a bias for genes with high mutual information scores and lower Pearson's correlation coefficients is evident. These gene pairs display nonlinear gene expression relationships and are dominated by transcription factors such as *TCF7L1* (Fig. 1B, highlighted in red), which is an important mutated gene in breast module 26.

Mutual information scores for gene expression levels, I, were used to construct symmetric, undirected, weighted, adjacency matrices, A, for which self connections were not allowed (i.e., the diagonal of the matrix is set to 0), and the connection strength between genes x and y is simply equivalent to the mutual information score over the expression levels of the genes, such that the elements of the I are defined as: $a_{xy} = a_{yx} = I(x;y)$. The mutual information scores were standardized so that the maximum mutual information score in each network was set to one. These mutual information scores were then used to define the weighted adjacency matrices, A, which were transformed to approximate an unweighted scale-free network topology of the type observed in other well-characterized biological systems (Zhang and Horvath 2005; Khanin and Wit 2006).

Cancer coexpression modules

The transformed adjacency matrices, A, were converted to distance matrices by replacing each value in the matrix by one minus the original value (i.e., the elements of the distance matrices were defined as: $a_{xy} = a_{yx} = 1-[I(x;y)/I(\max)]^s$. Where $I(\max)$ is the maximum mutual information score in the matrix (i.e., the standardization factor), and s is an integer used to transform the unweighted adjacency matrix to approximate the scale-free criteria (see Supplemental Methods for details).

These distance matrices were then subjected to hierarchical clustering with complete linkage. The distance and clustering methods used have demonstrated superior performance in similar contexts (Gibbons and Roth 2002). Finally, gene coexpression modules were defined by identifying closely connected subclusters using the Dynamic Tree Cut algorithm (Langfelder et al. 2008). To demonstrate the robustness of this approach, we randomly removed genes from the original data and analyzed the resultant modules for the existence of the original breast module 26 (see Supplemental Text). The approach gives the best results with a large gene set represented in the expression datasets.

This process resulted in 64 distinct breast cancer gene coexpression modules containing 10,379 genes (Supplemental Table 1), 137 colorectal cancer gene coexpression modules containing 11,531 genes (Supplemental Table 2), and 81 glioblastoma cancer gene coexpression modules containing 11,391 genes (Table 1; Supplemental Table 3). Most breast cancer modules overlapped significantly with a colorectal cancer or glioblastoma module, approximately half of the colorectal cancer modules overlapped significantly with a breast cancer module, approximately a third of colorectal cancer modules overlapped significantly with a glioblastoma module, about three-fourths of glioblastoma modules overlapped significantly with a breast cancer module, and a little less than half the glioblastoma modules overlapped significantly with a colorectal cancer module (as determined by the hypergeometric

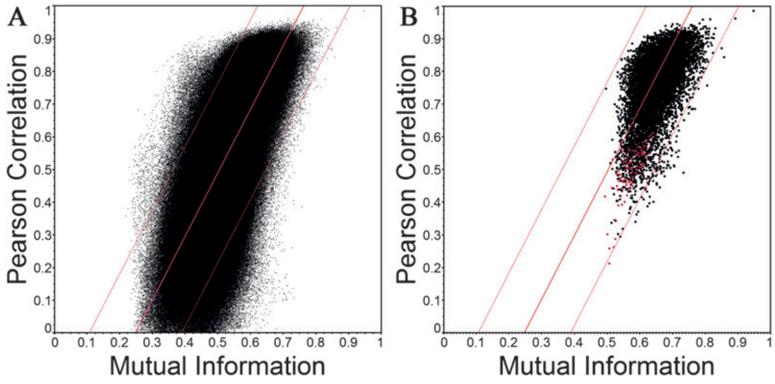

Figure 1. Mutual information vs. Pearson correlation. (*A*) Scatterplot of mutual information scores vs. Pearson's correlation coefficients for genes in breast module 26 vs. all other genes. The two scores are strongly correlated $R^2 = 0.53$. The *central* line is the linear fit with 95% confidence bands. (*B*) Scatterplot of mutual information scores vs. Pearson's correlation coefficients for genes in breast module 26 with only other genes in breast module 26. Scores involving *TCF7L1* are highlighted in red. Linear fit and 95% band of *A* is overlaid to demonstrate bias for high mutual information scores vs. Pearson's correlation coefficients.

Table 1. Summary of module characteristics

Tissue	No. of genes in modules	No. of modules	Median module size	Breast overlap	Colon overlap	Glioblastoma overlap
Breast	10,379	64	124		50 (78%)	50 (78%)
Colorectal	11,531	137	58	67 (49%)		45 (33%)
Glioblastoma	11,391	81	105	56 (69%)	36 (44%)	

Entries display the number of genes mapping to modules, the total number of modules per tissue, the median module size per tissue, and the number and percentage of overlapping modules per tissue pair.

distribution, threshold for significance corrected for multiple tests ≈ 5.0×10^{-6}; see Supplemental Methods for determination of significance thresholds) (Table 1). The relationships between the modules can be observed in Figure 2. Breast, colorectal cancer, and glioblastoma modules were subjected to two-way clustering based upon their probability of overlap as determined by the hypergeometric distribution, and are visualized as the negative log of the probability of overlap (Fig. 2). Distinct clusters of modules can be observed, and most modules from one tissue at least approach significance in overlap with a module from the opposing tissues. However, there exist unique modules in each tissue, suggesting an overall similarity in the gene coexpression modules in breast cancer and colorectal cancer, with a handful of tissue-specific modules. Overall, breast cancer and glioblastoma modules display the greatest degree of overlap (Table 1), a probable reflection of their common descent from the ectodermal germ layer as compared with the endodermal origin of colon tissue.

Identification of mutated coexpression modules

To identify coexpression modules containing a significant number of genes mutated in breast or colorectal cancer, we mapped the breast cancer, colon cancer, and glioblastoma mutated genes from the "Discovery Screens" by Wood et al. (2007) and Parsons et al. (2008) to the coexpression modules that we identified. The discovery screen by Wood et al. (2007) involved an initial genome-wide sequencing of 11 breast and 11 colorectal cancer samples, while the discovery screen by Parsons et al. (2008) involved an initial genome-wide sequencing of 22 glioblastoma samples (one hypermutated sample treated with temozolomide was removed from our analysis). The results of this sequencing effort were later used to identify what are referred to as frequently mutated, or "CAN," genes that were studied in later selective sequencing protocols in a larger set of additional tumor samples. By only focusing on genes mutated frequently within the "Discovery Screen," the investigators thus discarded potentially important cancer genes that are either mutated less frequently overall or simply happened to show a less-frequent mutation rate in the small sample at their disposal due to sampling error.

In our analyses, each mutated gene was counted only once within a module, despite the number of mutations identified in it. Thus, modules containing a single frequently mutated gene (for example, *TP53*) would not be considered further because of the presence of a single highly mutated gene. Our goal was to identify networks containing multiple, though rarely mutated genes important to tumorigenesis, rather than networks containing, e.g., a single gene that happened to be mutated frequently. We hypothesize that genes that mutated frequently play a stronger role in initiating tumorigenesis and are observed in all cancer cells of an individual tumor, whereas rarer mutations that cluster in gene modules are likely to be associated with processes involved in tumor progression, maintenance, and metastasis, and may be restricted to local populations of cancer cells within an individual tumor, though they are not necessarily excluded from involvement in tumor initiation. If these rarer mutations are involved in more "downstream" tumor progression events, they may also be more relevant to differential outcomes observed in cancer patients.

The distinction between what we are seeking to identify and what others have investigated is an important one and forms the main theme of our research, whereas previous analyses have been based primarily on the characterization of highly mutated, so called "CAN genes," (Chittenden et al. 2008), or focus on known pathways or direct interactions in order to identify very general tumorigenic processes (Hernández et al. 2007; Lin et al. 2007). On the other hand, we have focused on the reconstruction of cancer coexpression modules in order to identify specific rare driver mutations based upon their co-occurrence in these coexpression modules.

Mapping mutations to coexpression modules

Genes mutated in breast cancer, colorectal cancer, and glioblastoma were mapped to the reconstructed coexpression modules, and the significance of the number of mutated genes mapping to each module was evaluated by the hypergeometric distribution (Fig. 3). There was no significant trend for mutation enrichment within modules containing mutated genes with longer coding regions (see Supplemental Text). When breast cancer mutated genes were mapped to breast cancer modules (Fig. 3A, triangles), breast module 26 and 27 contained a significant number of mutated genes after correction for multiple testing (Fig. 3A, red triangles; also see Table 2) (*P*-values = 4.39×10^{-5} and 2.32×10^{-5}, respectively). The threshold for significance is denoted by the dashed line in Figure 2A (threshold for significance corrected for multiple tests = 1.98×10^{-4}). Breast cancer modules 26 and 27 do not significantly overlap with any colon cancer or glioblastoma module; thus, no colon cancer or glioblastoma modules were significantly enriched with mutations identified from the breast cancer samples (Fig. 3C,D, triangles). Breast cancer module 27 clusters with breast cancer module 26 (Fig. 2A,B), strongly suggesting that these modules are functionally related and contain genuine rare tumorigenic mutations. Furthermore, by independently mapping the breast CAN genes (i.e., the frequently mutated genes discussed in Wood et al. [2007] and Parsons et al. [2008]) to breast modules 26 and 27, we observe that a greater number of breast CAN genes than expected by chance reside in breast module 26 (Table 2, *P*-value = 2.32×10^{-5}, threshold for significance corrected for multiple tests = 0.006). Finally, mapping of colon CAN mutations to breast cancer modules reveals that breast module 26 is marginally enriched with colon CAN mutations Table 2, *P*-value = 0.005) and ranks fifth in enrichment of colon mutations after highly enriched breast modules 16, 33, and 32 (discussed below) (Fig. 3B, green circle; Table 2, *P*-value = 0.029). The probability that by random chance breast module 26 would be ranked numbers 1, 2, 5, and 2 in modules enriched with breast mutations, breast CAN, colon mutations, and colon CAN, respectively, is 7.0×10^{-6} by the rank product test (Breitling et al. 2004).

Breast modules 16 and 33 overlap significantly with colon module 6 and are discussed in that context below. Breast module

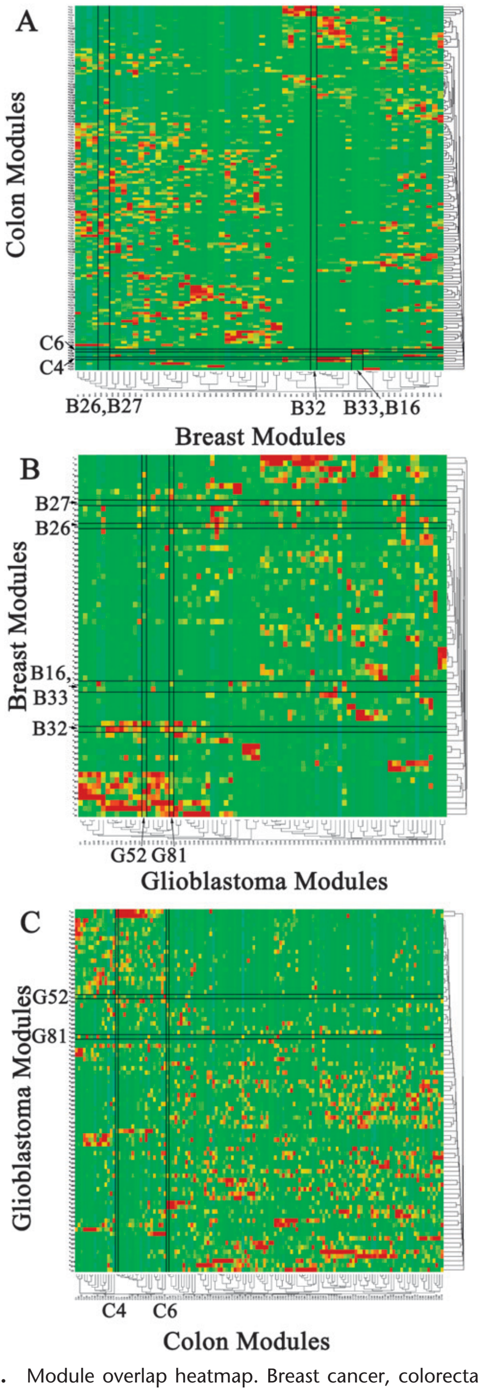

Figure 2. Module overlap heatmap. Breast cancer, colorectal cancer, and glioblastoma modules were subject to two-way clustering based upon their degree of overlap according to the hypergeometric distribution. (*A*) Colon modules are depicted on the vertical axis and breast modules are depicted on the horizontal axis. (*B*) Breast modules are depicted on the vertical axis and glioblastoma modules are depicted on the horizontal axis. (*C*) Glioblastoma modules are depicted on the vertical axis and colon modules are depicted on the horizontal axis. Modules of interest are labeled as B (breast), C (colon), or G (glioblastoma), followed by the module number. Coloration depicts the degree of overlap, where green indicates no overlap and red indicates the greatest degree of overlap for each module pair. Distinct clusters of overlapping modules can be observed. Note the close clustering of B26/B27 and B16/B33 in *A* and *B*. Note the significant overlap of B16/B33 with C6 in *A*.

32 is not significantly enriched with breast cancer mutations or breast CAN mutations; however, it is the most enriched module when glioblastoma mutations are mapped to breast modules (Fig. 3B, green diamond; Table 2, *P*-value = 0.0003) and ranks fourth when both colon mutations and colon CAN mutations are mapped to breast modules (Fig. 3B, green circle; Table 2, *P*-value = 0.029 [colon mutations], 0.02 [colon CAN]). The probability that by random chance breast module 32 would be ranked numbers 1, 4, and 4 in modules enriched with glioblastoma mutations, colon mutations and colon CAN, respectively, is 0.0008 by the rank product test.

When colon cancer mutated genes were mapped to colon modules (Fig. 3A, circles), colon modules 6 and 4 contained a significant number of mutated genes (Fig. 3A, red circles; Table 2, *P*-values = 8.1×10^{-6} and 1.8×10^{-4}, respectively). No additional evidence was available to support the association of colon module 4 with tumorigenesis. However, additional evidence strongly validates colon module 6's association with tumorigenesis. First, mapping of colon CAN genes to colon cancer coexpression modules indicated a marginally significant enrichment of colon CAN genes in colon module 6 (Table 2, *P*-value = 0.018). Colon module 6 was also the second most enriched module when mapping glioblastoma mutations to colon modules (Fig. 3C, green diamond; Table 2, *P*-value = 0.0015). Furthermore, colon module 6 significantly overlaps with breast cancer modules 16 and 33. When colorectal cancer mutants were mapped to breast cancer modules (Fig. 3B, circles), breast cancer module 16 is highly significant (Fig. 3B, yellow circle, *P*-value = 5.03×10^{-6}) and breast cancer module 33 is the next most enriched and marginally significant module (Fig. 3B, green circle, *P*-value 0.0039), as may be expected by their overlaps with colon module 6 (Fig. 1A). However, of the colorectal cancer mutated genes mapping to breast module 16, 48% (11 of 23 mutated genes) do not belong to colon cancer module 6, and of the colorectal cancer mutated genes mapping to breast cancer module 33, 64% (seven of 11 mutated genes) do not belong to colon module 6. Thus, the overlap of breast cancer modules 16 and 33 with colon cancer module 6 does not fully account for the clustering of colon cancer mutations within these modules and suggests that specific biological processes shared across these modules are disrupted by mutations during tumorigenesis.

The relationships shared by breast cancer modules 16 and 33 are further borne out by their close clustering in Figure 2, A and B. Additionally, mapping of colon cancer CAN genes to breast cancer modules 16 and 33 confirms a significant enrichment of colon cancer tumorigenic mutations within these modules (*P*-value = 0.0068 and 0.0004, respectively). Finally, and most strikingly, breast cancer module 33 is the next most enriched module (after the closely related significant breast cancer modules 26 and 27) when breast cancer mutants are mapped to breast cancer modules (Fig. 3A, green triangle; Table 2, *P*-value = 0.0036), and, in addition, is marginally enriched with breast cancer CAN genes (Table 2, *P*-value = 0.004) and glioblastoma mutations (Fig. 3B, green diamond; Table 2, *P*-value = 0.008). Thus, in two independent cancer mutation sets, derived from different tissues, a module corresponding to colon-cancer module 6 is enriched with cancer mutations, further suggesting that the biological processes associated with colon cancer module 6 are associated with tumorigenesis. The probability that, by random chance, breast module 33 would be ranked numbers 3, 6, 2, 1, and 2 in modules enriched with breast mutations, breast CAN, colon mutations, colon CAN, and glioblastoma mutations, respectively, is $<1.0 \times 10^{-6}$ by the rank product test.

Finally, when glioblastoma mutated genes were mapped to glioblastoma modules (Fig. 3A, diamonds), glioblastoma module 81

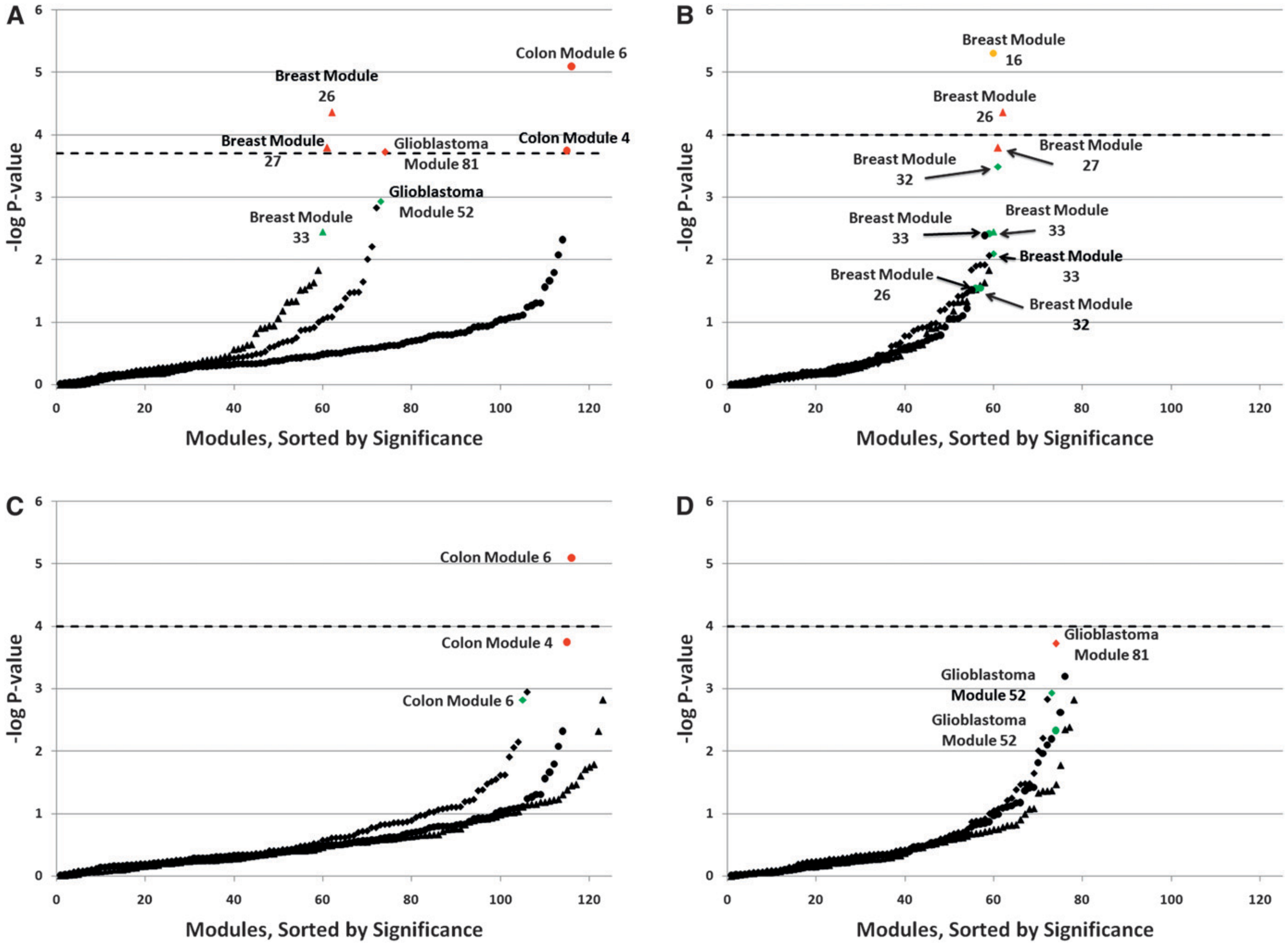

Figure 3. Enrichment of cancer mutations in breast and colorectal modules: The negative log *P*-value, as determined by the hypergeometric distribution, of the number of mutations mapping to each coexpression module, is plotted in order of significance. The threshold for statistical significance is denoted by the dashed line. (*A*) Mutations mapped to their corresponding tissue. (*B*) Breast modules; (*C*) colon modules; and (*D*) glioblastoma modules. In each panel, triangles represent modules mapped with breast mutations, circles represent modules mapped with colon mutations, and diamonds represent modules mapped with glioblastoma mutations. Red points are statistically significant for mutations mapped to their corresponding tissue, i.e., above the significance threshold in *A*. Yellow points are statistically significant for mutations mapped to other tissues. Green points are marginally significant modules of interest.

contained a significant number of mutated genes (Fig. 3A, red diamond; Table 2, *P*-value = 1.89×10^{-4}). Glioblastoma module 81 is the smallest glioblastoma module and does not significantly overlap with any breast or colon cancer module; thus, no colon cancer or glioblastoma modules were significantly enriched with mutations from glioblastoma samples. However, glioblastoma module 81 does contain a marginal enrichment of colon mutations and colon CAN mutations, ranking as modules 8 and 7, respectively. The probability that by random chance glioblastoma module 81 would be ranked numbers 1, 8, and 7 in modules enriched with glioblastoma mutations, colon cancer mutations, and colon CAN mutations, respectively, is 0.0026 by the rank product test. Note that there are only 20 glioblastoma CAN mutations mapping to any of our modules, and thus, do not provide striking results in any analyses.

The second most enriched module when glioblastoma mutations are mapped to glioblastoma modules is glioblastoma module 52 (Fig. 3A, green diamond; Table 2, *P*-value = 0.001). Glioblastoma module 52 also contains a marginal enrichment of colon mutations and colon CAN mutations (Fig. 3D, green circle; Table 2, *P*-value = 0.005 [colon mutations], 0.03 [colon CAN]), ranking as modules 3 and 8, respectively. The probability that by random chance glioblastoma module 52 would be ranked numbers 2, 3, and 8 in modules enriched with glioblastoma mutations, colon cancer mutations and colon CAN mutations, respectively, is 0.0021 by the rank product test.

Breast modules 26, colon module 6, and glioblastoma module 81 are clearly associated with tumorigenesis through straightforward statistical significance of the enrichment of mutations from their respective tissues, high rankings across independent tissue mutation datasets, and their relationships with other significant modules. Colon module 4 and breast module 27 are also clearly significant, though they are not supported by mutations from independent tissue mutation datasets. The marginally significant modules, breast modules 16, 33, 32, and glioblastoma module 52 are supported by their relationships with the strongly significant modules as well as their high enrichment rankings across independent tissue mutation datasets. Although these results lend credence to the involvement of all of these modules in tumorigenesis, we confine further discussion

Table 2. Candidate cancer modules

Module	Breast mutations	Breast CAN	Colon mutations	Colon CAN	Glioblastoma mutations	Glioblastoma CAN
Modules significant by mutations of the same tissue type						
Breast-26	**20 (4.39×10^{-5}) (1/62)**	**10 (2.32×10^{-5}) (2/37)**	*10 (0.029) (5/60)*	**4 (0.005) (2/34)**	9 (0.05) (11/61)	0 (1.00) (NA)
Breast-27	**18 (0.00016) (2/62)**	4 (0.07) (16/37)	4 (0.63) (43/60)	0 (1.00) (N/A)	1 (0.97) (59/61)	0 (1.00) (NA)
Colon-6	10 (0.85) (116/123)	1 (0.62) (54/56)	**26 (8.10×10^{-6}) (1/116)**	**5 (0.018) (4/56)**	**20 (0.0015) (2/106)**	0 (1.00) (NA)
Colon-4	14 (0.57) (92/123)	1 (0.65) (56/56)	**24 (1.8×10^{-4}) (2/116)**	2 (0.38) (54/56)	15 (0.08) (20/106)	0 (1.0) (NA)
Glio-81	2 (0.25) (28/78)	0 (1.00) (N/A)	*3 (0.038) (8/76)*	*1 (0.022) (7/46)*	**6 (1.89×10^{-4}) (1/74)**	0 (1.00) (NA)
Modules suggestive of enrichment by overall trends						
Breast-16	10 (0.66) (42/62)	2 (0.69) (29/37)	**23 (5.03×10^{-6}) (1/60)**	**5 (0.0068) (3/34)**	*13 (0.038) (9/61)*	0 (1.00) (NA)
Breast-33	**14 (0.0036) (3/62)**	**6 (0.004) (6/37)**	**11 (0.0039) (2/60)**	**5 (0.0004) (1/34)**	**10 (0.008) (2/61)**	1 (0.02) (8/16)
Breast-32	8 (0.27) (22/62)	0 (1.00) (N/A)	*9 (0.029) (4/60)*	*3 (0.02) (4/34)*	**13 (0.0003) (1/61)**	0 (1.00) (NA)
Glio-52	*9 (0.03) (5/78)*	1 (0.18) (25/48)	**9 (0.005) (3/76)**	*2 (0.03) (8/46)*	**10 (0.001) (2/74)**	1 (0.01) (4/18)

Entries include the number of mutations mapping to each module, the significance of mutation enrichment in each module, and the rank of the module out of the total number of modules containing mutations in each mutation category. Bold and underlined entries are significant, bold-only entries are marginally significant and high ranked, italicized entries are suggestive by rank and significance. Breast cancer modules 16 and 33 are closely related by clustering and overlap significantly with colon module 6. Breast cancer modules 26 and 27 are closely related by clustering. Threshold for significance for mutations mapped to their corresponding tissue is 1.98×10^{-4} and to other tissues is 9.92×10^{-5}. Threshold for significance for CAN mutations mapped to their corresponding tissue is 0.006 and to other tissues is 0.003.
NA, Not available.

to the functions targeted by mutations in the strongly significant modules: breast cancer module 26, colon cancer module 6, and glioblastoma module 81. A list of candidate cancer drivers mutated in each of the above modules is presented in the Supplemental Text.

Discussion

We performed gene ontology, literature, and interaction searches in order to characterize the molecular relationships between the mutated genes in breast cancer module 26 (Fig. 4), colon cancer module 6 (Fig. 5), and glioblastoma module 81 (Fig. 6). A full description of the interactions presented in Figures 4–6 is presented in the Supplemental Text. These figures are only meant to suggest how mutated genes within these modules are functionally connected based upon our best interpretation of available biological information. Importantly, coexpression modules should not be interpreted as coherent functional modules, though the genes contained within coexpression modules can be expected to contribute to similar biological processes due to their coregulation. Through this analysis, we identified Wnt/TGF-beta cross-talk, Wnt/VEGF signaling, and MAPK/focal adhesion kinase pathways as targets of rare driver mutations in breast cancer, colorectal cancer, and glioblastoma, respectively

Breast module 26 (Fig. 4) contains mutated genes, the majority of which are transcriptional/translational regulators that interact with SMAD3 and beta-catenin (CTNNB1), transcriptional mediators of the TGF-beta and Wnt signaling pathways. These signaling pathways can result in both tumor-suppressive and proliferative effects. TGF-beta is traditionally known for its tumor-suppressive effects, yet in later stages of tumor development the TGF-beta promotes tumor invasive processes (Massagué 2008). Beta-catenin (*CTNNB1*) is upregulated in ~60% of breast tumors, yet the activation of this pathway does not occur through mutation of beta-catenin

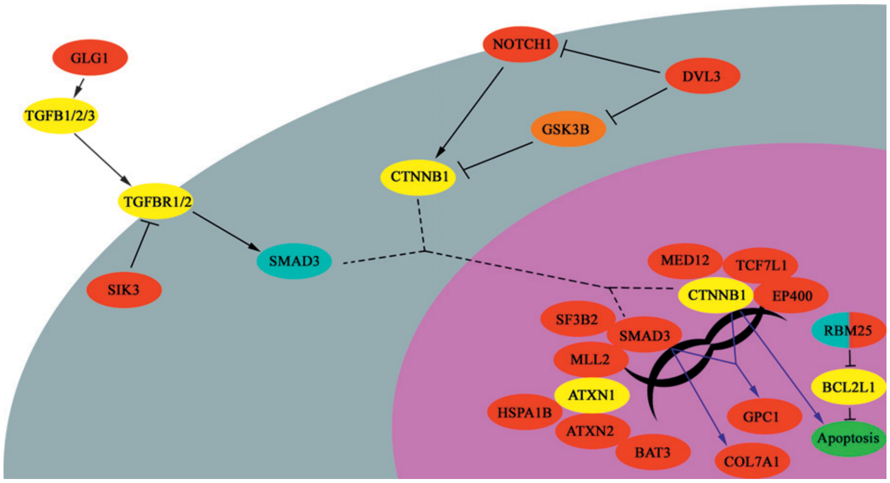

Figure 4. The relationship of genes mutated in breast module 26. (Red ovals) Genes mutated in breast cancer module 26; (aqua ovals) breast module 26 genes mutated in colon cancer; (orange ovals) the closely related breast module 27 genes; (yellow ovals) genes not present in breast cancer module 26. (Black lines or touching ovals) Functional protein relationships; (blue lines) transcriptional relationships. Dotted lines depict movement of SMAD3 and beta-catenin (CTNNB1) from the cytoplasm to nucleus.

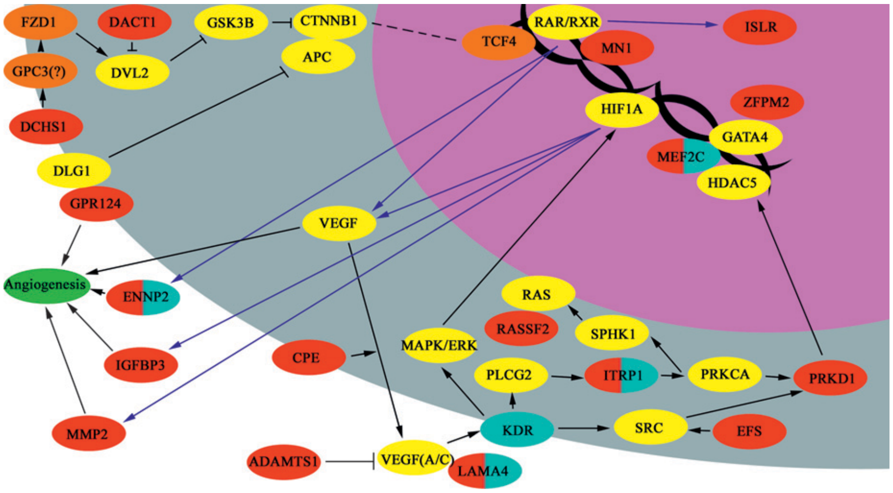

Figure 5. The relationship of genes mutated in colon cancer module 6. (Red oval) Genes mutated in colon cancer module 6; (aqua ovals) colon module 6 genes mutated in breast cancer or glioblastoma; (orange ovals) unmutated colon module 6 genes; (yellow ovals) genes not present in colon cancer module 6. (Black lines or touching ovals) Functional protein relationships; (blue lines) transcriptional relationships. Dotted lines depict movement of beta-catenin (CTNNB1) from the cytoplasm to nucleus.

(*CTNNB1*) or *APC*, which is the case in almost 100% of colon cancers (Lin et al. 2000). On the other hand, accumulation of beta-catenin (CTNNB1) in the cytoplasm can lead to apoptosis, which is mediated by BCL2L1 (also known as BCL-X) (Kim et al. 2000). Thus, we speculate that both pathways, as well as their cooperative effects, are manipulated by mutations in advanced tumor, so that the tumor-suppressive functions are inhibited while the tumor promoting activities are activated, or left intact. For example, mutations in *RBM25* would directly silence the tumor-suppressive signaling wing of the Wnt pathway, since it is involved in splicing of *BCL2L1* (also known as *BCL-X*) to its apoptotic form (Zhou et al. 2008).

Colon module 6 mutations appear to target induction of angiogenesis and VEGF secretion by the Wnt pathway, VEGF handling and secretion, as well as a VEGF autocrine hypoxia loop (Fig. 5). These alterations, including differences in extracellular matrix proteins, matrix metalloproteases, and angiogenic factors prime the tumor environment for blood vessel formation and metastasis. Mutations within the VEGF autocrine signaling loop target the PKC/SRC and RAS-mediated arms of KDR (also known as VEGFR2), signaling shutting down the tumor-suppressive MEF2C response to VEGF, which may favor the MAPK/ERK arm of the VEGF autocrine signaling loop, which in turn activates HIF1A (hypoxia-inducible factor 1, alpha subunit) (Calvani et al. 2008). This autocrine loop increases expression of additional pro-angiogenic factors. The notion that colon module 6 mutations are involved in metastasis is confirmed by mutational analysis of metastases versus primary colorectal tumors, where *ENPP2* and *PLCG2* mutations were found only in metastases (Jones et al. 2008a).

Glioblastoma module 81 is the smallest glioblastoma module, which appears to contain neuron-specific accessory genes to the focal adhesion kinase/mitogen-activated kinase pathways, which differentially mediate cellular adhesion versus proliferation and migration (Ha et al. 2008; Bigarella et al. 2009). These processes are differentially regulated by intracellular levels of effector molecules, such as calcium and retinoic acid, whose levels appear to be tweaked by mutations in ion channels and *RBP3*, a retinoic acid shuttle (Crowe et al. 2003; Papi et al. 2007). These mutations appear to favor motility and proliferation of glioblastoma cells by targeting neuron-specific adhesion mechanisms.

We believe a gene network reconstruction, strategy-based approach can successfully identify rare cancer driver mutations through enrichment of mutations within modules. The interplay of pathways described by our approach is unlikely to be detected by traditional pathway analysis approaches or a focus on frequently mutated genes. Our findings suggest that these rare mutations are involved in more peripheral elements of important tumorigenic signaling pathways, and we speculate that these rare mutations contribute to tumorigenesis by biasing the ultimate functional effects of these signaling pathways toward their tumor-promoting versus tumor-suppressive outcomes. Identifying cell lines containing these mutations, such as those used in the Wood et al. (2007) study, restoring the wild-type version of these mutated genes, and comparing the growth rates of the original versus "restored" cell lines in a variety of in vivo and in vitro contexts could be a means to verify the functional role of these mutants. Our approach should be described in the light of a few important caveats. The specific techniques used to reconstruct genetic networks can be altered to generate networks of different sizes, or reflecting different

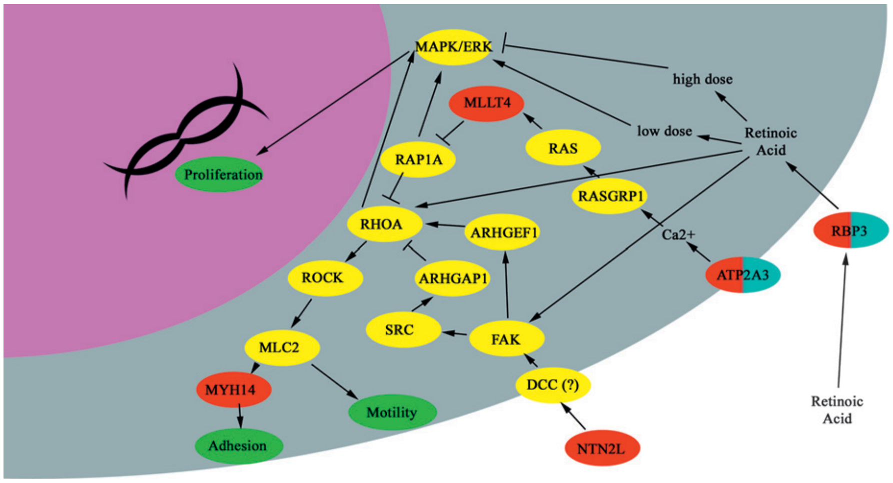

Figure 6. The relationship of genes mutated in glioblastoma module 81. (Red ovals) genes mutated in glioblastoma module 81; (aqua ovals) glioblastoma module 81 genes mutated in breast or colon cancer; (yellow ovals) genes not present in glioblastoma module 81. (Black lines) Functional protein relationships.

coexpression relationships, depending upon the investigators requirements and/or sample size and likelihood that module enrichment will be observed in different-sized modules. Additionally, our approach probably does not capture all of the secondary driver mutations, which may require either additional complementary systems biology approaches, or larger sample sizes to capture other mutation-enriched coexpression modules. Overall, we believe this approach shows tremendous promise for the identification of these rare tumorigenic driver mutations, which is a crucial task for upcoming large-scale cancer resequencing projects, as it is these more private mutations that may be driving intra-tumor heterogeneity, inter-patient heterogeneity, and ultimately altering response to therapeutic intervention.

Methods

Gene expression datasets from normal and cancerous breast and colorectal tissue were downloaded from the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/). Missing gene expression values were filled in as the average across the entire data set. To maximize the number of data sets available for breast, colorectal, and glioblastoma networks, we reconstructed networks based upon the Affymetrix Human Genome U133A platform (20,842 probes, 13,077 genes) using numerous datasets representing multiple perturbed states of normal and cancerous breast (466 samples in total) and colorectal (233 samples in total) and glial (463 samples in total) tissue (see Supplemental Methods for datasets used).

Mutual information scores quantifying relationships between the expression levels of genes within the normal and cancer tissue samples were calculated using the ARACNE algorithm with a *P*-value cutoff of 1×10^{-10} (Margolin et al. 2006). Adjacency matrices using the mutual information scores as its elements was converted to a distance matrix as described in the Results section (see Supplemental Methods for further details). Distance matrices derived from these adjacency matrices were subjected to hierarchical clustering with complete linkage using the *R* computational suite. Resulting trees from the cluster analyses were cut into subclusters using the Dynamic Tree Cut algorithm implemented in the cutreeHybrid approach in *R*. Probabalistic significance levels for overlapping clusters and the number of somatically mutated genes mapping to each cluster was calculated using the hypergeometric distribution in *R*. Gene Ontology enrichment of mutated genes within defined clusters of genes was determined using the Gene Ontology Tree Machine (Zhang et al. 2004). Protein–protein interactions and relationships between genes were investigated using a combination of Pathway Commons (http://www.pathwaycommons.org/), GeneCards (http://www.genecards.org), PubMed (http://www.pubmed.com), and Ali Baba (http://alibaba.informatik.hu-berlin.de).

Acknowledgments

N.J.S. and his laboratory are supported in part by the following research grants: The National Institute on Aging Longevity Consortium (U19 AG023122-01); The NIMH-funded Genetic Association Information Network Study of Bipolar Disorder National (1 R01 MH078151-01A1); National Institutes of Health (N01 MH22005, U01 DA024417-01, and P50 MH081755-01); Scripps Genomic Medicine and the Scripps Translational Sciences Institute Clinical Translational Science Award (U54 RR0252204-01). A.T. is a Scripps Genomic Medicine Dickinson Scholar.

References

Bigarella CL, Borges L, Costa FF, Saad ST. 2009. ARHGAP21 modulates FAK activity and impairs glioblastoma cell migration. *Biochim Biophys Acta* **1793:** 806–816.

Breitling R, Armengaud P, Amtmann A, Herzyk P. 2004. Rank products: A simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett* **573:** 83–92.

Butte AJ, Kohane IS. 2000. Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput* **5:** 415–429.

Calvani M, Trisciuoglio D, Bergamaschi C, Shoemaker RH, Melillo G. 2008. Differential involvement of vascular endothelial growth factor in the survival of hypoxic colon cancer cells. *Cancer Res* **68:** 285–291.

Cancer Genome Atlas Research Network. 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455:** 1061–1068.

Chittenden TW, Howe EA, Culhane AC, Sultana R, Taylor JM, Holmes C, Quackenbush J. 2008. Functional classification analysis of somatically mutated genes in human breast and colorectal cancers. *Genomics* **91:** 508–511.

Crowe DL, Kim R, Chandraratna RA. 2003. Retinoic acid differentially regulates cancer cell proliferation via dose-dependent modulation of the mitogen-activated protein kinase pathway. *Mol Cancer Res* **1:** 532–540.

Daub CO, Sonnhammer EL. 2008. Employing conservation of co-expression to improve functional inference. *BMC Syst Biol* **2:** 81. doi: 10.1186/1752-0509-2-81.

Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, et al. 2008. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455:** 1069–1075.

Gibbons FD, Roth FP. 2002. Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res* **12:** 1574–1581.

Ha VL, Bharti S, Inoue H, Vass WC, Campa F, Nie Z, deGramont A, Ward Y, Randazzo PA. 2008. ASAP3 is a focal adhesion-associated Arf GAP that functions in cell migration and invasion. *J Biol Chem* **283:** 14915–14926.

Hernández P, Solé X, Valls J, Moreno V, Capellá G, Urruticoechea A, Pujana MA. 2007. Integrative analysis of a cancer somatic mutome. *Mol Cancer* **6:** 13. doi: 10.1186/1476-4598-6-13.

Horvath S, Zhang B, Carlson M, Lu KV, Zhu S, Felciano RM, Laurance MF, Zhao W, Qi S, Chen Z, et al. 2006. Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. *Proc Natl Acad Sci* **103:** 17402–17407.

Jones S, Chen WD, Parmigiani G, Diehl F, Beerenwinkel N, Antal T, Traulsen A, Nowak MA, Siegel C, Velculescu VE, et al. 2008a. Comparative lesion sequencing provides insights into tumor evolution. *Proc Natl Acad Sci* **105:** 4283–4288.

Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, et al. 2008b. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321:** 1801–1806.

Khanin R, Wit E. 2006. How scale-free are biological networks. *J Comput Biol* **13:** 810–818.

Kim K, Pang KM, Evans M, Hay ED. 2000. Overexpression of beta-catenin induces apoptosis independent of its transactivation function with LEF-1 or the involvement of major G1 cell cycle regulators. *Mol Biol Cell* **11:** 3509–3523.

Langfelder P, Zhang B, Horvath S. 2008. Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R. *Bioinformatics* **24:** 719–720.

Lin SY, Xia W, Wang JC, Kwong KY, Spohn B, Wen Y, Pestell RG, Hung MC. 2000. Beta-catenin, a novel prognostic marker for breast cancer: Its roles in cyclin D1 expression and cancer progression. *Proc Natl Acad Sci* **97:** 4262–4266.

Lin J, Gan CM, Zhang X, Jones S, Sjöblom T, Wood LD, Parsons DW, Papadopoulos N, Kinzler KW, Vogelstein B, et al. 2007. A multidimensional analysis of genes mutated in breast and colorectal cancers. *Genome Res* **17:** 1304–1318.

Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A. 2006. ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7:** S7. doi: 10.1186/1471-2105-7-S1-S7.

Massagué J. 2008. TGFβ in cancer. *Cell* **134:** 215–230.

Oldham MC, Horvath S, Geschwind DH. 2006. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proc Natl Acad Sci* **103:** 17973–17978.

Papi A, Bartolini G, Ammar K, Guerra F, Ferreri AM, Rocchi P, Orlandi M. 2007. Inhibitory effects of retinoic acid and IIF on growth, migration and invasiveness in the U87MG human glioblastoma cell line. *Oncol Rep* **18:** 1015–1021.

Parsons DW, Jones S, Zhang X, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Siu IM, Gallia GL, et al. 2008. An integrated genomic analysis of human glioblastoma multiforme. *Science* **321:** 1807–1812.

Priness I, Maimon O, Ben-Gal I. 2007. Evaluation of gene-expression clustering via mutual information distance measure. *BMC Bioinformatics* **8:** 111. doi: 10.1186/1471-2105-8-111.

Torkamani A, Verkhivker G, Schork NJ. 2008. Cancer driver mutations in protein kinase genes. *Cancer Lett.* **281:** 117–127.

Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, et al. 2007. The genomic landscapes of human breast and colorectal cancers. *Science* **318:** 1108–1113.

Zhang B, Horvath S. 2005. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* **4:** Article 17. doi: 10.2202/1544-6115.1128.

Zhang B, Schmoyer D, Kirov S, Snoddy J. 2004. GOTree Machine (GOTM): A web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics* **5:** 16. doi: 10.1186/1471-2105-5-16.

Zhou A, Ou AC, Cho A, Benz EJ Jr, Huang SC. 2008. Novel splicing factor RBM25 modulates Bcl-x pre-mRNA 5′ splice site selection. *Mol Cell Biol* **28:** 5924–5936.

Received February 16, 2009; accepted in revised form June 18, 2009.